

Tesis Sony

by Fakultas Hukum

Submission date: 07-Jan-2026 08:03PM (UTC+0700)

Submission ID: 2697511150

File name: 20260106_Tesis.docx (5.68M)

Word count: 22674

Character count: 163020

Peningkatan Akurasi Deteksi Degradasi Model Klasifikasi Berbasis
Multi-Criteria Decision Model



Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer (M.Kom)

Nama: Sony Harianto
NIM: 14230030

⁷
**PROGRAM STUDI MAGISTER ILMU KOMPUTER (S2) FAKULTAS
TEKNOLOGI INFORMASI UNIVERSITAS NUSA MANDIRI
JAKARTA
2025**

DAFTAR ISI

DAFTAR ISI	i
BAB I	1
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Identifikasi Masalah	3
1.3 Rumusan Masalah.....	5
1.4 Tujuan Penelitian	6
1.5 Ruang Lingkup Penelitian.....	8
1.6 Manfaat Penelitian.....	8
1.7 Hipotesis Penelitian	10
1.8 Sistematika Penulisan.....	11
BAB II	13
TINJAUAN PUSTAKA	13
2.1 Model Klasifikasi Berbasis CNN	13
2.2 Degradasi Model dan Fenomena <i>Drift</i>	17
2.3 MLOps	20
2.4 Single-Metric Monitoring dan Keterbatasannya.....	23
2.5 Multi-Metrics Monitoring dan Stability Metrics.....	26
2.6 Composite Health Score dan Weighted Sum Model (WSM)	29
2.7 Standar Internasional dan AI Governance	33
2.8 Penelitian Terdahulu dan Research Gap	36
2.9 Posisi dan Kontribusi Penelitian.....	42
BAB III	45
METODOLOGI PENELITIAN	45
3.1 Pendekatan dan Desain Penelitian	45
3.2 Arsitektur Sistem Monitoring Model	47
3.3 Model dan Dataset Penelitian	50
3.4 Skenario Degradasi Data.....	53
3.5 Mekanisme Monitoring dan Metrik Evaluasi	56
3.6 Perhitungan Composite Health Score	59
3.7 Prosedur Eksperimen	63

3.8	Kriteria Evaluasi dan Analisis Perbandingan	66
3.9	Validitas Penelitian dan Keterbatasan Metodologi.....	69
17	BAB IV	72
	HASIL PENELITIAN DAN PEMBAHASAN	72
4.1	Gambaran Umum Pelaksanaan Eksperimen	72
4.2	Hasil Baseline Model pada Kondisi Normal.....	74
4.3	Analisis Degradasi Model Berdasarkan <i>Single-Metric Monitoring</i>	76
4.4	Analisis Degradasi Model Berdasarkan Multi-Metrics Monitoring ...	79
4.5	Hasil Perhitungan <i>Composite Health Score</i>	84
4.6	Perbandingan <i>Single-Metric</i> dan <i>Multi-Criteria Health Check</i>	88
4.7	Pembahasan Hasil dalam Konteks MLOps dan AI Governance	91
4.8	Ringkasan Temuan.....	95
	BAB V	97
	KESIMPULAN DAN REKOMENDASI	97
5.1	Kesimpulan	97
5.2	Kontribusi Penelitian	98
5.3	Keterbatasan Penelitian dan Rekomendasi Penelitian Selanjutnya ..	99
	DAFTAR PUSTAKA.....	i

DAFTAR GAMBAR

Gambar 1.1 Perbedaan Kondisi Model pada Fase Pelatihan dan Lingkungan Produksi.....	1
Gambar 1.2 Ilustrasi Degradasi Model Klasifikasi CNN akibat Penurunan Kualitas Data Input	3
Gambar 2.1 Arsitektur Umum Model Klasifikasi Berbasis CNN (MobileNetV3).....	14
Gambar 2.2 Arsitektur CNN sebagai Model Probabilistik.....	15
Gambar 2.3 Ilustrasi Degradasi Model Klasifikasi CNN akibat Penurunan Kualitas Data Input.....	16
Gambar 2.4 Ilustrasi Model Degradation dan Data Drift pada Sistem Machine Learning.....	18
Gambar 2.5 Alur Umum MLOps.....	21
Gambar 2.6 Perbandingan Pendekatan Single-Metric dan Multi-Metric dalam Monitoring Model.....	24
Gambar 2.7 Kerangka Multi-Metrik untuk Monitoring Kesehatan Model.....	26
Gambar 2.8 Taxonomy MCDM.....	30
Gambar 2.9 Tahapan Perhitungan WSM.....	31
Gambar 2.10 Mapping Monitoring terhadap ISO/IEC.....	35
Gambar 3.1 Desain Penelitian Eksperimental untuk Deteksi Degradasi Model.....	45
Gambar 3.2 Arsitektur Sistem Monitoring Model dalam Lingkungan MLOps.....	47
Gambar 3.3 Skema Monitoring Model Berbasis Batch.....	50
Gambar 3.4 Contoh Skenario Degradasi Data Visual.....	56
Gambar 3.5 Single vs Multi-Metrics Monitoring.....	57
Gambar 3.6 Proses Normalisasi dan Agregasi Metrik menggunakan WSM.....	60
Gambar 3.7 Prosedur Eksperimen.....	60
Gambar 4.1 Perubahan Confidence Ratio pada Berbagai Skenario Degradasi (Single Matriks).....	79

Gambar 4.2 Perubahan PSI terhadap degradasi data.....	81
Gambar 4.3 Perubahan KL Divergence terhadap degradasi data.....	82
Gambar 4.4 Perubahan Latency Inference akibat Degradasi Data (Multi metrics).83	
Gambar 4.5 Perubahan Composite Health Score (D_prod) per Batch.....	85

DAFTAR TABEL

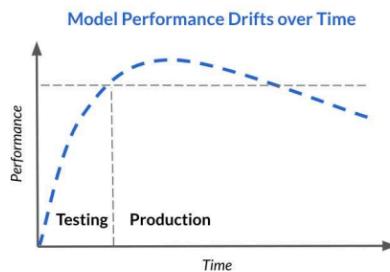
Tabel 1.1 Fokus Rumusan Masalah.....	5
Tabel 1.2 Pemetaan Tujuan Penelitian dan Output yang Diharapkan.....	7
Tabel 1.3 Ruang Lingkup dan Batasan Penelitian.....	8
Tabel 2.1 Karakteristik Model CNN dalam Konteks Deployment Produksi.....	16
Tabel 2.2 Perbandingan Jenis Drift dan Dampaknya terhadap Model.....	17
Tabel 2.3 Perbandingan Jenis Perbandingan Single vs Multi-Metric.....	25
Tabel 2.4 Definisi dan Fungsi Metrik.....	29
Tabel 2.5 Skema Bobot WSM.....	31
Tabel 2.6 Kesesuaian Penelitian dengan Standar.....	35
Tabel 2.7 Penelitian Terdahulu.....	36
Tabel 2.8 Posisi Penelitian terhadap State-of-the-Art.....	43
Tabel 3.1 Komponen Arsitektur Monitoring.....	47
Tabel 3.2 Spesifikasi Model.....	50
Tabel 3.3 Ringkasan Dataset.....	52
Tabel 3.4 Daftar Skenario Degradasi.....	53
Tabel 3.5 Definisi Metrik Monitoring.....	59
Tabel 3.6 Skema Bobot Composite Score.....	61
Tabel 3.7 Klasifikasi Composite Health Score dan Tindakan Operasional.....	62
Tabel 3.8 Tahapan Eksperimen.....	66
Tabel 3.9 Kriteria Evaluasi.....	67
Tabel 3.10 Validitas dan Keterbatasan.....	69
Tabel 4.1 Performa Baseline Model (Tanpa Degradasi).....	73
Tabel 4.2 Status Model Berdasarkan Single-Metric Monitoring.....	79
Tabel 4.3 Ringkasan Nilai Multi-Metrics per Skenario.....	84
Tabel 4.4 Nilai Composite Health Score dan Status Model.....	88
Tabel 4.5 Perbandingan Waktu Deteksi Degradasi.....	89
Tabel 4.6 Implikasi Monitoring terhadap Keputusan Rollback.....	91
Tabel 4.7 Ringkasan Temuan Utama.....	96

33
BAB I
PENDAHULUAN

1.1 Latar Belakang

Penerapan model *machine learning* pada lingkungan produksi telah menjadi praktik umum dalam berbagai sistem cerdas modern, khususnya pada sistem klasifikasi berbasis *computer vision*. Model yang telah mencapai performa tinggi pada tahap pelatihan tidak selalu mempertahankan kinerja yang sama setelah diimplementasikan di lingkungan nyata. Perubahan karakteristik data masukan, kondisi lingkungan, serta dinamika operasional sistem dapat menyebabkan penurunan performa model secara bertahap maupun tiba-tiba [1][3].

Fenomena penurunan performa model setelah deployment dikenal sebagai degradasi model, yang umumnya dipicu oleh perubahan distribusi data atau *drift* [21][27] [29]. Degradasi ini dapat muncul dalam berbagai bentuk, seperti perubahan pencahayaan, tingkat kebisingan (*noise*), kualitas sensor, atau variasi konteks yang tidak sepenuhnya terwakili pada data pelatihan. Dalam sistem produksi, degradasi model yang tidak terdeteksi secara dini berpotensi menghasilkan keputusan yang tidak akurat dan berdampak langsung pada kualitas layanan maupun risiko operasional [5][6][31].



Gambar 1.1 Perbedaan Kondisi Model pada Fase Pelatihan dan Lingkungan Produksi

Sumber: <https://blog.devgenius.io/unsimply-model-decay-aa3689975ada>

Dalam praktik *Machine Learning Operations* (MLOps), proses pemantauan kinerja model umumnya masih mengandalkan satu atau dua metrik evaluasi utama, seperti akurasi atau *F1-score*. Pendekatan *single-metric* monitoring ini memiliki keterbatasan karena tidak mampu merepresentasikan kondisi kesehatan model secara menyeluruh, terutama ketika degradasi terjadi secara gradual atau bersifat parsial. Model dapat menunjukkan nilai akurasi yang relatif stabil, sementara metrik lain seperti stabilitas prediksi, latensi, atau konsistensi distribusi data telah mengalami penurunan [8][9][10].

Keterbatasan pendekatan tersebut berdampak pada keterlambatan pengambilan keputusan, khususnya dalam konteks *rollback* model. Proses *rollback* yang lambat atau tidak tepat waktu dapat meningkatkan risiko kesalahan prediksi dalam jangka waktu yang signifikan sebelum tindakan korektif dilakukan. Oleh karena itu, diperlukan mekanisme pemantauan yang mampu memberikan gambaran kondisi model secara lebih komprehensif dan sensitif terhadap berbagai aspek [8][11][12] [14] [19].

Pendekatan *multi-criteria health check* menawarkan alternatif dengan mengombinasikan beberapa metrik evaluasi, baik yang bersifat performa maupun stabilitas, ke dalam satu indikator komposit yang merepresentasikan kondisi kesehatan model secara keseluruhan. Dengan adanya indikator kesehatan komposit ini, proses deteksi degradasi dan pengambilan keputusan *rollback* diharapkan dapat dilakukan secara lebih cepat, objektif, dan terukur dalam kerangka MLOps yang sistematis. Berdasarkan latar belakang tersebut, penelitian ini difokuskan pada pengembangan dan evaluasi pendekatan *multi-criteria health check* untuk mendeteksi degradasi model klasifikasi secara lebih efektif, serta menganalisis kontribusinya terhadap percepatan proses *rollback* pada lingkungan MLOps.



Gambar 1.2 Ilustrasi Degradasi Model Pasca Deployment dalam Lingkungan
MLOps

Sumber: Penulis, 2025

1.2 Identifikasi Masalah

Berdasarkan latar belakang tersebut, permasalahan utama dalam pengelolaan model klasifikasi berbasis CNN tidak terletak pada kemampuan model dalam mencapai performa awal, melainkan pada keberlanjutan dan stabilitas perilaku

model setelah deployment. Data operasional yang bersifat dinamis menyebabkan model rentan mengalami degradasi meskipun tidak terjadi perubahan pada arsitektur maupun parameter model [5][6][31].

Praktik monitoring model yang masih banyak mengandalkan metrik performa tradisional, seperti akurasi atau *F1-score*, memiliki keterbatasan mendasar. Metrik-metrik tersebut bersifat diskrit dan bergantung pada ketersediaan label *ground truth*, sehingga sering kali tidak mampu mendeteksi perubahan awal pada perilaku internal model. Akibatnya, degradasi baru teridentifikasi setelah dampaknya signifikan terhadap kualitas sistem [12][31]. *Confidence Ratio* menawarkan alternatif *single-metric* yang lebih sensitif karena merepresentasikan tingkat keyakinan probabilistik model. Namun, metrik ini tetap hanya mencerminkan satu aspek kesehatan model dan tidak menangkap perubahan distribusi probabilitas output, pergeseran proporsi kelas prediksi, maupun kondisi operasional sistem [12][14].

Di sisi lain, degradasi model juga ditandai oleh perubahan distribusi probabilitas output dan *class shift*, yang dapat diukur melalui stability metrics seperti ¹³ *Population Stability Index (PSI)* dan *Kullback–Leibler Divergence (KL Divergence)*. Meskipun demikian, metrik-metrik tersebut umumnya dipantau secara terpisah dan belum diintegrasikan ke dalam satu indikator komposit yang mudah diinterpretasikan untuk mendukung pengambilan keputusan operasional [11][12]. Studi-studi sebelumnya menunjukkan bahwa perubahan distribusi probabilitas sering kali menjadi sinyal awal terjadinya degradasi model sebelum dampaknya terlihat pada metrik performa tradisional [5][31]. Namun demikian, metrik stabilitas ini masih jarang diintegrasikan secara sistematis dalam mekanisme monitoring model pada lingkungan MLOps.

Permasalahan lain yang turut muncul adalah adanya *class shift*, yaitu perubahan proporsi prediksi antar kelas yang dapat mengindikasikan pergeseran pola data operasional atau munculnya bias baru pada model [21]. Perubahan ini sering kali tidak terdeteksi apabila sistem monitoring hanya berfokus pada nilai agregat seperti akurasi atau *Confidence Ratio*. Di sisi lain, aspek operasional seperti latensi inferensi juga sering dipantau secara terpisah, padahal peningkatan latensi

dapat menjadi indikasi awal bahwa model menghadapi input yang semakin kompleks atau tidak sesuai dengan distribusi pelatihan [8][19].

Lebih lanjut, meskipun standar internasional seperti ISO/IEC 23053 dan ISO/IEC 5338 telah menekankan pentingnya pemantauan sistem AI secara multidimensional—mencakup performa, stabilitas, dan aspek operasional—implementasi teknis dari rekomendasi tersebut masih terbatas dalam praktik [9][10]. Dengan demikian, terdapat kesenjangan antara kompleksitas perilaku model pembelajaran mesin di lingkungan produksi dan pendekatan monitoring yang saat ini banyak diterapkan. Kesenjangan tersebut mencakup keterbatasan pendekatan *single-metric*, kurangnya integrasi metrik stabilitas dan operasional, serta belum tersedianya indikator kesehatan model yang holistik dan siap digunakan sebagai dasar pengambilan keputusan rollback dalam pipeline MLOps.

8.1.3 Rumusan Masalah

Berdasarkan latar belakang dan identifikasi masalah yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

Bagaimana pendekatan *multi-criteria health check* berbasis *composite score* yang mengintegrasikan ¹³ *Population Stability Index* (PSI), *Kullback–Leibler Divergence* (KL Divergence), *Class Shift*, *Confidence Ratio*, dan *Latency* mampu mendeteksi degradasi model klasifikasi berbasis CNN secara lebih cepat dan sensitif dibandingkan pendekatan *single-metric* monitoring yang hanya menggunakan *Confidence Ratio* dalam lingkungan MLOps?

Rumusan masalah ini menekankan dua aspek utama, yaitu integrasi multi-metrik sebagai representasi kesehatan model secara holistik dan kecepatan deteksi degradasi sebagai faktor kunci dalam mendukung kesiapan serta efektivitas proses rollback. Dengan demikian, penelitian ini diarahkan untuk tidak hanya mengevaluasi performa model secara statis, tetapi juga menganalisis dinamika perilaku model dalam lingkungan operasional yang realistik.

Tabel 1.1 Fokus Rumusan Masalah

Aspek	Single-Metric	Multi-Criteria
Dimensi	Tunggal	Multidimensi
Sensitivitas	Terbatas	Tinggi

Kesiapan Rollback	Rendah	Tinggi
-------------------	--------	--------

⁴² 1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini adalah mengembangkan dan mengevaluasi pendekatan *multi-criteria health check* berbasis *composite score* yang mampu mendeteksi degradasi model klasifikasi lebih cepat dan lebih sensitif dibandingkan pendekatan *single-metric* monitoring yang hanya menggunakan *Confidence Ratio* dalam lingkungan MLOps. Tujuan utama ini menekankan aspek kecepatan deteksi degradasi sebagai faktor kunci dalam mendukung kesiapan dan efektivitas proses rollback model, sebagaimana direkomendasikan dalam praktik MLOps dan standar internasional [9][10][15].

Untuk mencapai tujuan utama tersebut, penelitian ini memiliki beberapa tujuan khusus yang saling terkait. Pertama, penelitian ini bertujuan untuk menganalisis perilaku degradasi model klasifikasi berbasis CNN ketika menghadapi berbagai skenario penurunan kualitas data visual yang merepresentasikan kondisi dunia nyata, seperti blur, penurunan pencahayaan, resolusi rendah, dan noise sensorik. Analisis ini dilakukan untuk memahami bagaimana perubahan kualitas input memengaruhi distribusi probabilitas output, tingkat kepercayaan prediksi, dan stabilitas perilaku model secara keseluruhan [6][21][27].

Kedua, penelitian ini bertujuan untuk mengevaluasi efektivitas *Confidence Ratio* sebagai pendekatan *single-metric* monitoring dalam mendeteksi degradasi model. *Confidence Ratio* dipilih karena mampu merepresentasikan tingkat keyakinan model terhadap prediksi yang dihasilkan dan telah ditunjukkan dalam berbagai studi sebagai indikator awal penurunan kualitas model [12][14]. Evaluasi ini dilakukan untuk mengukur sejauh mana *Confidence Ratio* mampu memberikan sinyal degradasi secara dini, serta mengidentifikasi keterbatasannya ketika digunakan sebagai satu-satunya indikator kesehatan model.

Ketiga, penelitian ini bertujuan untuk mengembangkan pendekatan *multi-metrics health check* dengan mengintegrasikan metrik performa, stabilitas distribusi, dan aspek operasional. Metrik yang digunakan dalam penelitian ini meliputi PSI, KL Divergence, Class Shift, Confidence Ratio, dan Latency. Integrasi

metrik-metrik tersebut bertujuan untuk merepresentasikan kesehatan model secara holistik, sesuai dengan rekomendasi evaluasi multidimensional dalam sistem AI [9][10][11][12].

Keempat, penelitian ini bertujuan untuk membangun *composite health score* menggunakan pendekatan *Multi-Criteria Decision Making* (MCDM), khususnya *Weighted Sum Model* (WSM). Pendekatan ini digunakan untuk menggabungkan berbagai metrik heterogen ke dalam satu nilai komposit yang mudah diinterpretasikan dan dapat digunakan sebagai dasar pengambilan keputusan operasional [13]. Tujuan ini mencakup penentuan skema normalisasi metrik, pembobotan kriteria, serta perhitungan skor komposit yang konsisten dan dapat direplikasi.

Kelima, penelitian ini bertujuan untuk membandingkan kecepatan deteksi degradasi antara pendekatan *single-metric monitoring* dan *multi-criteria health check*. Perbandingan dilakukan dengan mengamati titik waktu atau batch pertama yang menunjukkan indikasi degradasi pada masing-masing pendekatan. Analisis ini bertujuan untuk memberikan bukti nyata mengenai apakah pendekatan multi-kriteria benar-benar mampu mendeteksi degradasi lebih awal dibandingkan penggunaan *Confidence Ratio* sebagai indikator tunggal [5][31].

Terakhir, penelitian ini bertujuan untuk memberikan rekomendasi konseptual dan teknis bagi pengembangan sistem monitoring model yang lebih adaptif dalam pipeline MLOps, khususnya dalam mendukung pengambilan keputusan *rollback* model. Rekomendasi ini diharapkan dapat menjadi referensi bagi praktisi dan peneliti dalam merancang mekanisme monitoring yang lebih selaras dengan kompleksitas perilaku model AI di lingkungan produksi serta dengan prinsip AI governance dan operational resilience [8][9][10].

Tabel 1.2 Pemetaan Tujuan Penelitian dan Output yang Diharapkan

No	Tujuan	Output	Relevansi
1	Analisis degradasi CNN	Pola perubahan metrik	Teoretis
2	Evaluasi single-metric	Baseline	Komparatif
3	Pengembangan multi-criteria	Composite score	Metodologis
4	Analisis rollback	Waktu deteksi	Operasional

³⁴ 1.5 Ruang Lingkup Penelitian

Penelitian ini difokuskan pada analisis degradasi model klasifikasi berbasis CNN dengan arsitektur MobileNetV3 pada skenario klasifikasi dua kelas. Degradasi yang dianalisis dibatasi pada penurunan kualitas visual input, meliputi blur, penurunan pencahayaan, resolusi rendah, dan *noise*. Penelitian ini tidak membahas strategi korektif lanjutan seperti retraining atau online learning, melainkan berfokus pada mekanisme monitoring dan deteksi degradasi untuk mendukung keputusan rollback model. Penelitian ini dibatasi pada degradasi yang disebabkan oleh penurunan kualitas visual input, yaitu blur, penurunan pencahayaan, resolusi rendah, dan noise injection, yang merepresentasikan kondisi umum pada sistem computer vision berbasis kamera [21][27].

Tabel 1.3 Ruang Lingkup dan Batasan Penelitian

Aspek	Batasan
Jenis Model	CNN berbasis image classification
Lingkungan	Simulasi production
Jenis Degradasi	Blur, lighting, noise, compression
Fokus	Monitoring & rollback decision

1.6 Manfaat Penelitian

Secara teoritis, penelitian ini memberikan kontribusi terhadap pengembangan kajian mengenai degradasi model dan monitoring model pembelajaran mesin berbasis *computer vision*. Selama ini, sebagian besar penelitian mengenai CNN berfokus pada peningkatan performa model pada fase pelatihan dan pengujian, sementara perilaku model setelah *deployment* masih relatif kurang mendapat perhatian [1][3][5]. Penelitian ini memperluas perspektif tersebut dengan menempatkan degradasi model sebagai fenomena operasional yang bersifat multidimensi, bukan sekadar penurunan akurasi. Dengan mengintegrasikan stability metrics seperti PSI dan KL Divergence, penelitian ini memperkaya literatur mengenai pendekatan deteksi degradasi berbasis distribusi probabilitas, yang telah terbukti lebih sensitif terhadap perubahan awal perilaku model [11][12][31].

Selain itu, penelitian ini memberikan kontribusi metodologis melalui penerapan pendekatan MCDM, khususnya WSM, dalam konteks monitoring

kesehatan model pembelajaran mesin. Meskipun WSM telah banyak digunakan dalam pengambilan keputusan multikriteria pada berbagai domain, penerapannya sebagai mekanisme penggabungan metrik monitoring model dalam lingkungan MLOps masih sangat terbatas [13]. Dengan demikian, penelitian ini membuka peluang pengembangan lebih lanjut mengenai integrasi metode MCDM dalam sistem evaluasi dan pengendalian kualitas model AI.

Dari sisi praktis, penelitian ini memberikan manfaat langsung bagi organisasi yang mengoperasikan model pembelajaran mesin dalam lingkungan produksi. Pendekatan *multi-criteria health check* yang diusulkan memungkinkan deteksi degradasi model secara lebih dini dibandingkan pendekatan *single-metric* monitoring yang hanya mengandalkan satu indikator, seperti *Confidence Ratio*. Deteksi dini ini sangat penting untuk mengurangi risiko operasional akibat model yang terus berjalan dalam kondisi tidak stabil, seperti peningkatan kesalahan prediksi, penurunan kualitas layanan, atau gangguan proses bisnis [8][19]. Dengan adanya *composite health score*, tim operasional tidak perlu memantau banyak metrik secara terpisah, melainkan cukup mengamati satu indikator komposit yang lebih mudah diinterpretasikan dan dapat digunakan sebagai dasar pengambilan keputusan.⁶

Dari sisi strategis dan tata kelola AI, penelitian ini mendukung penerapan prinsip pemantauan sistem AI yang lebih transparan, terstruktur, dan bertanggung jawab. Standar internasional seperti ISO/IEC 23053 dan ISO/IEC 5338 menekankan bahwa sistem AI harus dipantau secara berkelanjutan dengan mempertimbangkan berbagai dimensi performa dan risiko operasional [9][10]. Pendekatan multi-criteria health check yang dikembangkan dalam penelitian ini dapat menjadi contoh implementasi teknis dari rekomendasi standar tersebut, sehingga membantu organisasi dalam memenuhi tuntutan tata kelola AI, kepatuhan regulasi, serta akuntabilitas penggunaan sistem AI.

Secara keseluruhan, manfaat penelitian ini tidak hanya terbatas pada pengembangan konsep dan metode deteksi degradasi model, tetapi juga mencakup peningkatan praktik operasional dan tata kelola sistem AI di lingkungan produksi. Dengan menghubungkan pendekatan teknis monitoring model dengan kebutuhan operasional dan standar internasional, penelitian ini diharapkan dapat menjadi

referensi bagi peneliti, praktisi, maupun pengambil kebijakan dalam merancang sistem MLOps yang lebih adaptif, andal, dan berkelanjutan.

1.7 Hipotesis Penelitian⁴

Berdasarkan latar belakang, identifikasi masalah, rumusan masalah, serta tujuan penelitian yang telah diuraikan pada subbab sebelumnya, penelitian ini dirancang untuk menguji efektivitas pendekatan monitoring model dalam mendeteksi degradasi model klasifikasi berbasis CNN pada lingkungan MLOps. Secara khusus, penelitian ini membandingkan pendekatan single-metric monitoring berbasis *Confidence Ratio* dengan pendekatan *multi-criteria health check* yang mengintegrasikan berbagai metrik evaluasi ke dalam sebuah *composite health score*.

Pendekatan *single-metric* monitoring diposisikan sebagai *baseline* yang merepresentasikan praktik monitoring sederhana yang umum digunakan dalam sistem produksi. Sementara itu, pendekatan *multi-criteria health check* dirancang untuk menangkap degradasi model yang bersifat multidimensional, dengan mengombinasikan metrik stabilitas distribusi, kepercayaan prediksi, perubahan proporsi kelas, serta aspek operasional sistem. Dengan demikian, penelitian ini berangkat dari asumsi bahwa degradasi model tidak selalu tercermin secara langsung pada satu indikator tunggal, melainkan muncul melalui perubahan bertahap pada berbagai dimensi perilaku model.

²⁴ Berdasarkan kerangka pemikiran tersebut, hipotesis yang diajukan dalam penelitian ini adalah sebagai berikut:

H1: Pendekatan *multi-criteria health check* berbasis *composite health score* mampu mendeteksi degradasi model klasifikasi berbasis CNN secara lebih cepat dibandingkan pendekatan *single-metric* monitoring yang hanya menggunakan *Confidence Ratio* dalam lingkungan MLOps.

H2: Pendekatan *multi-criteria health check* menghasilkan sinyal deteksi degradasi model yang lebih konsisten dan stabil dibandingkan pendekatan *single-metric* monitoring, khususnya pada skenario degradasi data visual yang bersifat gradual.

H3: Integrasi metrik stabilitas distribusi, metrik kepercayaan prediksi, dan metrik operasional dalam *composite health score* memberikan dasar pengambilan

keputusan rollback model yang lebih objektif dibandingkan penggunaan satu metrik monitoring secara terpisah.

Hipotesis-hipotesis tersebut diuji melalui eksperimen terkontrol dengan menerapkan berbagai skenario degradasi data visual pada model klasifikasi berbasis CNN yang telah dideploy. Pengujian hipotesis difokuskan pada analisis temporal perubahan metrik monitoring dan *composite health score*, serta pada perbandingan kecepatan dan konsistensi deteksi degradasi antara pendekatan *single-metric* dan *multi-criteria*. Dengan demikian, pengujian hipotesis dalam penelitian ini tidak hanya bertujuan untuk membuktikan keunggulan pendekatan multi-kriteria secara teknis, tetapi juga untuk menilai relevansinya dalam mendukung kesiapan pengambilan keputusan operasional, khususnya rollback model, dalam pipeline MLOps.

1.8 Sistematika Penulisan

Penulisan tesis ini disusun secara sistematis ke dalam lima bab yang saling berkaitan dan membentuk satu kesatuan yang utuh. Sistematika penulisan ini dirancang untuk mengarahkan pembaca dalam memahami alur pemikiran penelitian, mulai dari perumusan permasalahan hingga penyampaian kesimpulan dan rekomendasi berdasarkan hasil penelitian yang telah dilakukan.¹⁸

Bab I merupakan pendahuluan yang menyajikan gambaran umum mengenai latar belakang penelitian, identifikasi masalah, rumusan masalah, tujuan penelitian, ruang lingkup penelitian, serta manfaat penelitian. Bab ini berfungsi sebagai landasan konseptual yang menjelaskan urgensi penelitian terkait degradasi model klasifikasi berbasis CNN dan pentingnya pendekatan *multi-criteria health check* dalam lingkungan MLOps.¹⁹

Bab II merupakan tinjauan pustaka yang membahas landasan teori dan kajian penelitian terdahulu yang relevan dengan topik penelitian. Pembahasan pada bab ini mencakup konsep model klasifikasi berbasis CNN, fenomena degradasi model dan berbagai bentuk *drift*, prinsip-prinsip MLOps, pendekatan monitoring model berbasis multi-metrik, teori MCDM dengan fokus pada WSM, serta standar internasional yang terkait dengan pengelolaan dan pemantauan sistem AI. Selain

itu, bab ini juga memuat analisis *research gap* dan posisi penelitian terhadap *state-of-the-art*.

Bab III membahas metodologi penelitian yang digunakan untuk menjawab rumusan masalah. Bab ini menjelaskan pendekatan penelitian, desain arsitektur sistem monitoring, rancangan eksperimen, skenario degradasi data, instrumen penelitian, serta tahapan pelaksanaan penelitian. Selain itu, bab ini juga menguraikan metrik evaluasi yang digunakan, mekanisme perhitungan *composite health score*, serta metode perbandingan antara pendekatan *single-metric* monitoring dan *multi-criteria health check*.

³² Bab IV menyajikan hasil penelitian dan pembahasan. Pada bab ini ditampilkan hasil eksperimen yang diperoleh dari pengujian model pada berbagai skenario degradasi, analisis perubahan metrik monitoring pada setiap batch, serta perhitungan *composite score*. Pembahasan dilakukan dengan menekankan perbandingan kecepatan deteksi degradasi antara pendekatan *single-metric Confidence Ratio* dan pendekatan *multi-metrics composite health score*, serta implikasinya terhadap kesiapan dan efektivitas proses *rollback* model dalam lingkungan MLOps.

⁴³ Bab V merupakan penutup yang berisi kesimpulan dan rekomendasi. Bab ini merangkum temuan utama penelitian, menjawab rumusan masalah dan tujuan penelitian, serta menegaskan kontribusi teoritis dan praktis dari penelitian yang dilakukan. Selain itu, bab ini juga menyajikan keterbatasan penelitian serta rekomendasi untuk pengembangan penelitian selanjutnya, khususnya terkait perluasan pendekatan monitoring model dan integrasinya ke dalam sistem MLOps yang lebih komprehensif.

30
BAB II
TINJAUAN PUSTAKA

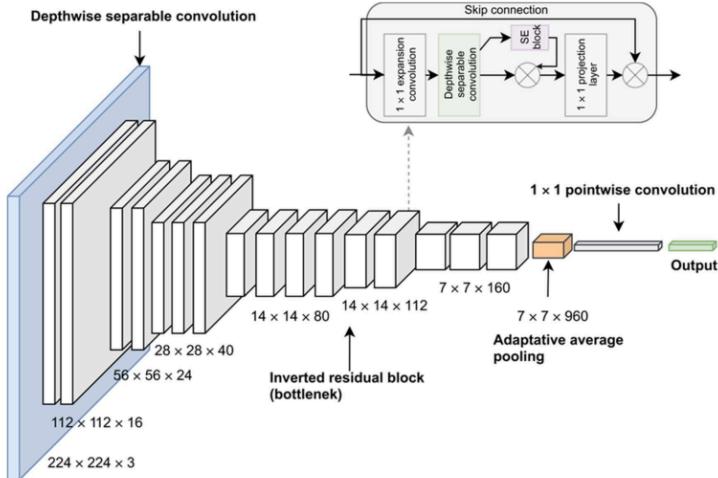
2.1 Model Klasifikasi Berbasis CNN

CNN merupakan arsitektur pembelajaran mendalam yang dirancang untuk memproses data visual berdimensi spasial melalui mekanisme ekstraksi fitur hierarkis. CNN memanfaatkan operasi konvolusi untuk menangkap pola lokal pada citra, yang kemudian dikombinasikan secara bertahap pada lapisan yang lebih dalam untuk merepresentasikan fitur tingkat tinggi. Karakteristik ini menjadikan CNN sangat efektif dalam tugas klasifikasi visual dan banyak diadopsi pada sistem computer vision modern.

2.1.1 Arsitektur Dasar CNN dan Mekanisme Ekstraksi Fitur

¹⁴ CNN terdiri dari beberapa komponen utama, yaitu *convolutional layer*, *pooling layer*, dan *fully connected layer*. *Convolutional layer* berfungsi mengekstraksi fitur lokal melalui operasi konvolusi menggunakan kernel, sementara *pooling layer* berfungsi melakukan reduksi dimensi untuk meningkatkan ketahanan model terhadap variasi spasial [1][3]. Lapisan *fully connected* kemudian memetakan fitur-fitur yang telah diekstraksi ke ruang keputusan untuk menghasilkan prediksi akhir.

Proses ekstraksi fitur dalam CNN bersifat hierarkis. Lapisan awal cenderung menangkap fitur rendah seperti tepi dan tekstur, sedangkan lapisan yang lebih dalam menangkap fitur tingkat tinggi seperti bentuk objek dan struktur semantik. Karakteristik ini menjadikan CNN sangat efektif untuk tugas klasifikasi visual, tetapi sekaligus membuatnya sensitif terhadap perubahan distribusi data input, khususnya perubahan kualitas visual yang tidak terwakili dalam data pelatihan [21][27].



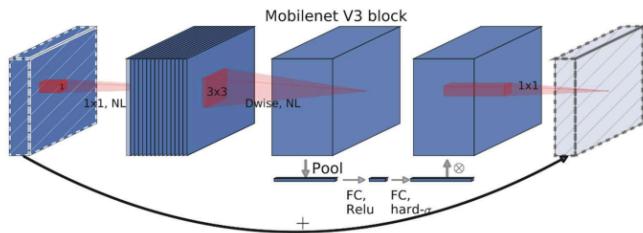
Gambar 2.1 Arsitektur Umum Model Klasifikasi Berbasis CNN (MobileNetV3)

Sumber: https://www.researchgate.net/figure/The-MobileNetV3-architecture-and-its-core-components_fig4_375462137

2.1.2 CNN sebagai Model Probabilistik

Dalam tugas klasifikasi, CNN umumnya menggunakan fungsi aktivasi softmax pada lapisan output untuk menghasilkan distribusi probabilitas terhadap setiap kelas. Distribusi probabilitas ini merepresentasikan tingkat keyakinan model terhadap setiap kemungkinan kelas, bukan sekadar keputusan kelas akhir [12]. Oleh karena itu, keluaran CNN secara inheren bersifat probabilistik. Sifat probabilistik ini memiliki implikasi penting dalam konteks monitoring model. Perubahan pada distribusi probabilitas output dapat terjadi meskipun prediksi kelas akhir masih tetap sama. Dengan kata lain, model dapat mempertahankan akurasi yang relatif stabil, sementara tingkat keyakinan prediksinya menurun secara bertahap. Fenomena ini sering kali menjadi indikasi awal terjadinya degradasi model yang tidak dapat ditangkap oleh metrik performa diskrit seperti akurasi atau F1-score [31]. Dalam penelitian ini, karakteristik probabilistik CNN menjadi dasar pemilihan Confidence Ratio sebagai representasi *single-metric* monitoring, serta penggunaan

metrik stabilitas distribusi seperti PSI dan KL *Divergence* untuk memantau perubahan perilaku model secara lebih mendalam.



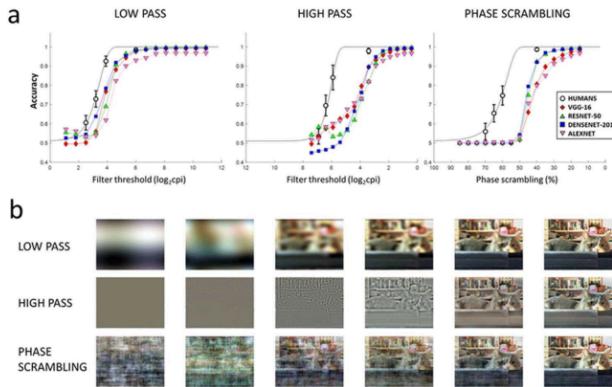
Gambar 2.2 Arsitektur CNN sebagai Model Probabilistik

Sumber: <https://pytorch.org/blog/torchvision-mobilenet-v3-implementation>

2.1.3 Sensitivitas CNN terhadap Perubahan Kualitas Data

Meskipun CNN memiliki kemampuan generalisasi yang kuat, berbagai penelitian menunjukkan bahwa performa dan stabilitas CNN sangat dipengaruhi oleh kualitas data input. Perubahan pencahayaan, *blur*, resolusi rendah, dan gangguan *noise* dapat menyebabkan pergeseran distribusi fitur yang diekstraksi oleh lapisan awal CNN, yang kemudian terpropagasi ke lapisan-lapisan berikutnya [21][27][29].

Sensitivitas ini menyebabkan CNN rentan mengalami degradasi perilaku ketika dioperasikan dalam lingkungan produksi yang dinamis. Perubahan kualitas visual yang bersifat gradual sering kali tidak langsung menurunkan akurasi, tetapi menyebabkan penurunan tingkat keyakinan model dan perubahan distribusi probabilitas output. Kondisi ini menegaskan bahwa evaluasi performa berbasis satu metrik tidak cukup untuk merepresentasikan kesehatan model secara menyeluruh.



Gambar 2.3 Ilustrasi Degradasi Model Klasifikasi CNN akibat Penurunan
Kualitas Data Input

Sumber: https://www.researchgate.net/figure/a-Degradation-accuracy-function-for-each-CNN-The-straight-black-line-represents-the_fig1_352742146

2.1.4 Implikasi CNN terhadap Monitoring dan Degradasi Model

Karakteristik CNN sebagai model hierarkis memiliki implikasi langsung terhadap strategi monitoring model. Karena degradasi model sering kali tercermin terlebih dahulu pada perubahan distribusi probabilitas dan tingkat keyakinan prediksi, maka pendekatan monitoring yang efektif harus mampu menangkap dinamika tersebut [12][31]. Dalam konteks ini, CNN menuntut pendekatan *monitoring* yang tidak hanya berfokus pada hasil akhir prediksi, tetapi juga pada stabilitas perilaku internal model. Hal ini menjadi dasar bagi penggunaan metrik stabilitas distribusi dan integrasi multi-metrik dalam composite health score yang diusulkan dalam penelitian ini. Dengan memahami karakteristik dasar CNN, dapat disimpulkan bahwa degradasi model merupakan fenomena yang bersifat gradual dan multidimensi, sehingga memerlukan pendekatan monitoring yang lebih komprehensif dibandingkan evaluasi performa konvensional.

Tabel 2.1 Karakteristik Model CNN dalam Konteks Deployment Produksi

Aspek	Karakteristik	Relevansi Monitoring
Feature extraction	Bertingkat	Sensitif distribusi

Confidence output	Probabilistik	Bisa dianalisis
Latency	Variatif	Dampak SLA
Black-box nature	Rendah interpretabilitas	Butuh monitoring

2.2 Degradasi Model dan Fenomena *Drift*

Degradasi model merujuk pada penurunan kualitas perilaku model setelah deployment akibat ketidaksesuaian antara data pelatihan dan data operasional yang diterima model seiring waktu. Berbeda dengan permasalahan underfitting atau overfitting yang terjadi pada fase pelatihan, degradasi model merupakan fenomena pasca-deployment yang dipicu oleh perubahan eksternal pada lingkungan data.

Tabel 2.2 Perbandingan Jenis Drift dan Dampaknya terhadap Model

Jenis Drift	Penyebab	Dampak
Data Drift	Perubahan distribusi input	<i>Confidence</i> menurun
Concept Drift	Perubahan relasi input-output	Akurasi turun
Population Drift	Perubahan proporsi kelas	Bias prediksi

2.2.1 Definisi dan Karakteristik Degradasi Model

Degradasi model (model *degradation* atau model *decay*) didefinisikan sebagai penurunan kualitas perilaku model setelah deployment akibat ketidaksesuaian antara data pelatihan dan data operasional yang diterima model seiring waktu [5][31]. Berbeda dengan *underfitting* atau *overfitting* yang terjadi pada fase pelatihan, degradasi model merupakan fenomena pasca-deployment yang dipicu oleh perubahan eksternal pada lingkungan data.

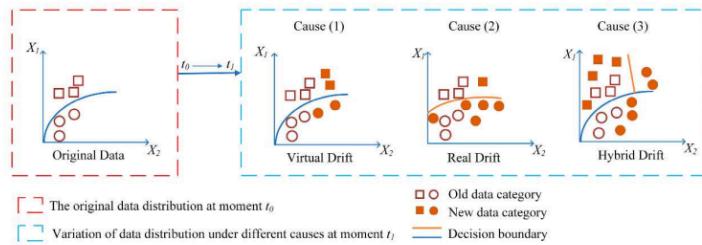
Salah satu karakteristik utama degradasi model adalah sifatnya yang gradual dan tidak selalu langsung tercermin pada metrik performa tradisional. Model dapat mempertahankan akurasi yang relatif stabil dalam periode tertentu, sementara distribusi probabilitas output dan tingkat keyakinan prediksi telah mengalami perubahan yang signifikan. Kondisi ini menyebabkan degradasi model sering kali luput dari deteksi apabila sistem monitoring hanya mengandalkan satu metrik performa diskrit [12][31].

2.2.2 Model Degradation dan Data Drift

Fenomena degradasi model umumnya berkaitan erat dengan berbagai bentuk *drift*, yaitu perubahan distribusi data atau hubungan antara data dan label. Secara umum, *drift* dapat diklasifikasikan ke dalam beberapa jenis utama. Data *drift* terjadi ketika distribusi data input berubah, sementara hubungan antara input dan label tetap relatif stabil. Perubahan ini dapat disebabkan oleh faktor lingkungan, perubahan perilaku pengguna, atau degradasi kualitas sensor [6][17]. Dalam konteks computer vision, data *drift* sering muncul dalam bentuk perubahan pencahayaan, sudut pandang kamera, atau kualitas citra.

Concept drift terjadi ketika hubungan antara data input dan label target mengalami perubahan. Pada kondisi ini, pola yang sebelumnya dipelajari oleh model tidak lagi valid, sehingga prediksi model menjadi kurang akurat meskipun distribusi input tampak serupa [17]. *Concept drift* lebih sulit dideteksi karena memerlukan pemahaman terhadap perubahan semantik pada data.

Selain itu, terdapat *population drift* atau *prior probability shift*, yaitu perubahan proporsi kelas dalam data operasional. Perubahan ini dapat menyebabkan model menghasilkan distribusi prediksi yang bias terhadap kelas tertentu, meskipun akurasi agregat belum menunjukkan penurunan yang signifikan [21]. Dalam penelitian ini, fenomena ini direpresentasikan melalui metrik class shift.



Gambar 2.4 Ilustrasi Model *Degradation* dan Data *Drift* pada Sistem *Machine Learning*

Sumber: <https://www.mdpi.com/2076-3417/13/11/6515>

2.2.3 Degradasi Model pada Sistem *Computer Vision*

Pada sistem *computer vision*, degradasi model sering kali dipicu oleh penurunan kualitas visual input. Faktor-faktor seperti blur akibat pergerakan kamera, perubahan intensitas pencahayaan, resolusi rendah akibat kompresi, serta gangguan *noise* sensorik dapat menyebabkan perubahan distribusi fitur yang diekstraksi oleh CNN [21][27]. Penelitian terdahulu menunjukkan bahwa CNN sangat sensitif terhadap perubahan kualitas visual, bahkan pada tingkat degradasi yang relatif ringan [27][29]. Perubahan tersebut dapat menyebabkan pergeseran distribusi probabilitas output sebelum penurunan akurasi terlihat secara signifikan. Hal ini menegaskan bahwa degradasi model pada sistem computer vision sering kali bersifat subtil dan tersembunyi, sehingga memerlukan pendekatan monitoring yang mampu menangkap perubahan perilaku internal model.

2.2.4 Keterbatasan Deteksi Degradasi Berbasis Performa

Pendekatan deteksi degradasi yang hanya mengandalkan metrik performa seperti akurasi atau *F1-score* memiliki keterbatasan mendasar. Metrik performa bersifat diskrit dan bergantung pada ketersediaan label *ground truth*, yang dalam banyak sistem produksi tidak selalu tersedia secara *real-time* [12][31]. Selain itu, penurunan performa sering kali merupakan indikator terlambat (*lagging indicator*), muncul setelah degradasi model terjadi dalam skala yang signifikan. Kondisi ini menyebabkan organisasi sering kali menyadari adanya degradasi model setelah dampaknya terasa pada kualitas layanan atau proses bisnis. Oleh karena itu, diperlukan pendekatan monitoring yang mampu mendekripsi degradasi secara lebih dini dengan memanfaatkan informasi distribusi probabilitas output dan karakteristik operasional model.

2.2.5 Implikasi *Drift* terhadap Strategi Monitoring Model

Fenomena *drift* dan degradasi model memiliki implikasi langsung terhadap strategi monitoring yang diterapkan dalam sistem MLOps. Karena degradasi sering kali tercermin terlebih dahulu pada perubahan distribusi probabilitas output dan tingkat keyakinan prediksi, maka pendekatan monitoring harus mampu mengukur perubahan tersebut secara kuantitatif [11][12].

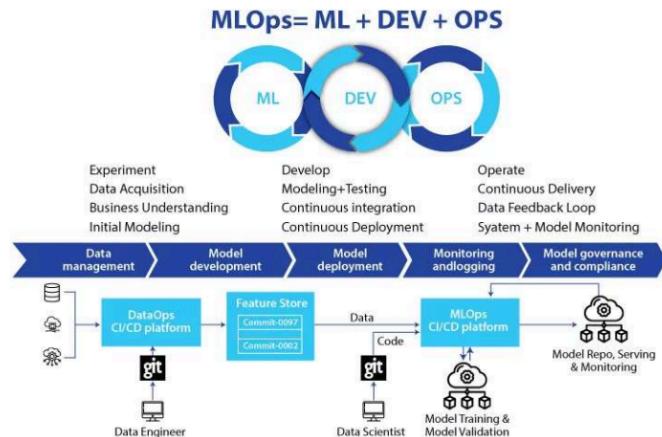
Metrik stabilitas distribusi seperti PSI dan KL *Divergence* memungkinkan pengukuran pergeseran distribusi output model secara eksplisit. Sementara itu,

pengukuran class shift memberikan indikasi perubahan proporsi prediksi antar kelas yang dapat mengindikasikan bias atau perubahan pola data operasional [21]. Integrasi metrik-metrik ini dengan indikator kepercayaan prediksi dan aspek operasional seperti latensi inferensi memungkinkan analisis degradasi model secara lebih holistik.

Dengan demikian, degradasi model dan fenomena *drift* tidak dapat dipahami sebagai permasalahan satu dimensi. Pendekatan monitoring yang efektif harus mempertimbangkan berbagai aspek perilaku model secara simultan. Pemahaman ini menjadi landasan konseptual bagi pengembangan pendekatan *multi-criteria health check* berbasis *composite score* yang diusulkan dalam penelitian ini, serta menjadi penghubung logis antara kajian degradasi model dan pembahasan mengenai monitoring multi-metrik.

2.3 MLOps

MLOps merupakan pendekatan sistematis yang mengintegrasikan praktik *software engineering*, *DevOps*, dan pembelajaran mesin untuk mengelola siklus hidup model secara berkelanjutan dalam lingkungan produksi. MLOps muncul sebagai respons terhadap meningkatnya kompleksitas sistem pembelajaran mesin, di mana model tidak lagi bersifat statis, melainkan beroperasi dalam ekosistem data yang dinamis dan terus berubah. Dalam konteks ini, MLOps tidak hanya berfokus pada proses deployment model, tetapi juga pada pemantauan, pemeliharaan, dan mitigasi risiko operasional akibat degradasi model [8][9].



Gambar 2.5 Alur Umum MLOps

Sumber: <https://www.igmguru.com/blog/machine-learning-operations-mlops-overview-definition-and-architecture>

2.3.1 Siklus Hidup Model dalam MLOps

Siklus hidup model dalam MLOps mencakup beberapa tahapan utama, yaitu pengumpulan dan persiapan data, pelatihan model, evaluasi, *deployment*, *monitoring*, serta pemeliharaan atau pemulihan model. Berbeda dengan pengembangan perangkat lunak konvensional, model pembelajaran mesin sangat bergantung pada data, sehingga kualitas dan karakteristik data operasional memiliki pengaruh langsung terhadap perilaku model setelah deployment [8].

Pada fase *deployment*, model mulai berinteraksi dengan data dunia nyata yang sering kali memiliki karakteristik berbeda dari data pelatihan. Oleh karena itu, fase monitoring menjadi krusial untuk memastikan bahwa model tetap beroperasi dalam batas performa dan stabilitas yang dapat diterima. Monitoring yang efektif memungkinkan deteksi dini terhadap degradasi model dan memfasilitasi tindakan mitigasi sebelum dampaknya meluas ke proses bisnis atau layanan publik.

2.3.2 Monitoring Model sebagai Komponen Kritis MLOps

Monitoring model merupakan salah satu komponen paling kritis dalam MLOps karena berfungsi sebagai mekanisme pengawasan berkelanjutan terhadap perilaku model. Tujuan utama monitoring bukan hanya untuk mendeteksi penurunan performa, tetapi juga untuk mengidentifikasi perubahan perilaku model yang berpotensi menimbulkan risiko operasional [10][15].

Dalam praktik, monitoring model mencakup berbagai aspek, termasuk performa prediksi, stabilitas distribusi data dan output, serta metrik operasional seperti latensi dan konsumsi sumber daya. Namun, banyak implementasi MLOps masih membatasi monitoring pada satu atau dua metrik performa, sehingga gagal menangkap dinamika perilaku model secara menyeluruh. Kondisi ini meningkatkan risiko bahwa degradasi model baru terdeteksi setelah sistem mengalami dampak yang signifikan.

2.3.3 Risiko Operasional akibat Degradasi Model

Degradasi model yang tidak terdeteksi secara tepat waktu dapat menimbulkan berbagai risiko operasional. Pada sistem yang menghasilkan keputusan otomatis, seperti sistem pengawasan lalu lintas atau keamanan publik, degradasi model dapat menyebabkan kesalahan deteksi, peningkatan *false decision*, dan penurunan kepercayaan pengguna terhadap sistem AI [19]. Selain itu, degradasi model juga dapat berdampak pada efisiensi sistem, misalnya melalui peningkatan latensi inferensi atau penggunaan sumber daya komputasi yang tidak optimal.

Dalam konteks MLOps, risiko operasional akibat degradasi model tidak hanya bersifat teknis, tetapi juga dapat berdampak pada aspek bisnis dan reputasi organisasi. Oleh karena itu, MLOps menekankan pentingnya mekanisme monitoring yang mampu memberikan sinyal peringatan dini (*early warning system*) terhadap degradasi model, sehingga tindakan mitigasi dapat dilakukan secara proaktif.

2.3.4 *Rollback* sebagai Strategi Mitigasi dalam MLOps

Salah satu strategi mitigasi yang umum digunakan dalam MLOps adalah *rollback* model, yaitu mengembalikan sistem ke versi model sebelumnya yang lebih stabil ketika degradasi terdeteksi. *Rollback* dipilih sebagai strategi awal karena

relatif cepat, berisiko rendah, dan tidak memerlukan data tambahan seperti retraining [15]. Efektivitas *rollback* sangat bergantung pada kemampuan sistem monitoring dalam mendeteksi degradasi model secara dini dan andal. Jika degradasi baru terdeteksi setelah dampaknya signifikan, maka *rollback* menjadi kurang efektif dan dapat menimbulkan gangguan layanan. Oleh karena itu, MLOps menuntut pendekatan monitoring yang tidak hanya akurat, tetapi juga sensitif terhadap perubahan awal perilaku model.

Dalam konteks penelitian ini, kebutuhan akan mekanisme *rollback* yang cepat dan objektif menjadi landasan utama pengembangan pendekatan multi-criteria health check. Dengan mengintegrasikan berbagai metrik performa, stabilitas, dan operasional ke dalam satu indikator komposit, diharapkan sistem monitoring dapat memberikan sinyal degradasi yang lebih dini dan mendukung pengambilan keputusan *rollback* secara lebih efektif.

2.4 *Single-Metric Monitoring* dan Keterbatasannya

Pendekatan *single-metric* monitoring merupakan praktik yang masih banyak digunakan dalam sistem pembelajaran mesin di lingkungan produksi. Pendekatan ini mengandalkan satu metrik utama sebagai indikator kondisi model, dengan asumsi bahwa metrik tersebut cukup merepresentasikan performa dan stabilitas model secara keseluruhan. Meskipun sederhana dan mudah diimplementasikan, pendekatan ini memiliki keterbatasan mendasar ketika dihadapkan pada kompleksitas perilaku model setelah *deployment*.

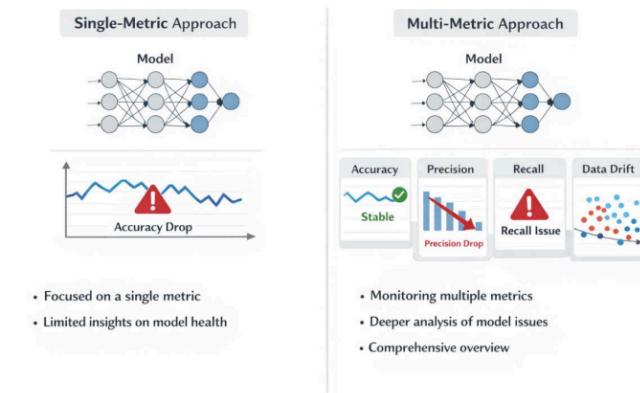
2.4.1 Konsep *Single-Metric* Monitoring dalam Praktik

Dalam praktik umum, *single-metric* monitoring biasanya menggunakan metrik performa seperti akurasi, *precision*, *recall*, atau *F1-score* untuk mengevaluasi kondisi model. Metrik-metrik ini bersifat diskrit dan bergantung pada ketersediaan label *ground truth* untuk menghitung kesesuaian antara prediksi model dan nilai aktual [12]. Pada fase pelatihan dan pengujian terkontrol, metrik tersebut efektif untuk membandingkan berbagai model dan mengukur performa secara objektif.

Namun, dalam lingkungan produksi, ketersediaan label *ground truth* sering kali terbatas atau tertunda, sehingga monitoring berbasis metrik performa menjadi

kurang praktis. Selain itu, metrik performa hanya mengukur hasil akhir prediksi dan tidak memberikan informasi mengenai dinamika internal model, seperti tingkat keyakinan prediksi atau perubahan distribusi probabilitas output.

Perbandingan Pendekatan Single-Metric dan Multi-Metric dalam Monitoring Model



Gambar 2.6 Perbandingan Pendekatan Single-Metric dan Multi-Metric dalam Monitoring Model

Sumber: Penulis, 2025

2.4.2 Akurasi sebagai Indikator yang Bersifat *Lagging*

Akurasi merupakan salah satu metrik yang paling umum digunakan dalam evaluasi model klasifikasi. Namun, sejumlah penelitian menunjukkan bahwa akurasi cenderung bersifat sebagai lagging *indicator* dalam konteks monitoring pasca-deployment [31]. Penurunan akurasi sering kali baru terlihat setelah degradasi model terjadi dalam skala yang cukup besar. Dalam banyak kasus, model dapat mempertahankan nilai akurasi yang relatif stabil meskipun distribusi probabilitas output telah mengalami pergeseran signifikan. Kondisi ini menimbulkan ilusi bahwa model masih berfungsi dengan baik, padahal tingkat keyakinan prediksi dan stabilitas perilaku model telah menurun. Oleh karena itu, ketergantungan pada akurasi sebagai satu-satunya indikator monitoring berpotensi menunda deteksi degradasi dan meningkatkan risiko operasional.

2.4.3 *Confidence Ratio* sebagai *Single-Metric* Alternatif

Sebagai respons terhadap keterbatasan metrik performa tradisional, beberapa penelitian mengusulkan penggunaan metrik berbasis probabilitas untuk monitoring model. *Confidence Ratio* merupakan salah satu metrik yang merepresentasikan tingkat keyakinan model terhadap prediksi kelas yang dihasilkan, yang dihitung berdasarkan distribusi probabilitas output model [12][14].

Confidence Ratio memiliki keunggulan dibandingkan akurasi karena mampu menangkap perubahan tingkat keyakinan model meskipun prediksi kelas akhir masih benar. Penurunan *Confidence Ratio* sering kali muncul lebih awal dibandingkan penurunan akurasi, sehingga dapat berfungsi sebagai indikator awal degradasi model. Dalam penelitian ini, *Confidence Ratio* digunakan sebagai representasi *single-metric* monitoring yang lebih sensitif dan relevan dibandingkan metrik performa diskrit.

Meskipun demikian, *Confidence Ratio* tetap merupakan metrik tunggal yang hanya merepresentasikan satu aspek kesehatan model, yaitu tingkat keyakinan prediksi. *Confidence Ratio* tidak secara eksplisit mengukur perubahan distribusi probabilitas antar waktu, pergeseran proporsi kelas prediksi, maupun aspek operasional seperti latensi inferensi.

2.4.4 Keterbatasan Fundamental Pendekatan *Single-Metric*

Keterbatasan utama pendekatan *single-metric* monitoring terletak pada ketidakmampuannya untuk merepresentasikan degradasi model yang bersifat multidimensi. Degradasi model tidak hanya ditandai oleh penurunan performa atau kepercayaan prediksi, tetapi juga oleh perubahan distribusi probabilitas output, *class shift*, dan kondisi operasional sistem [11][12][21].

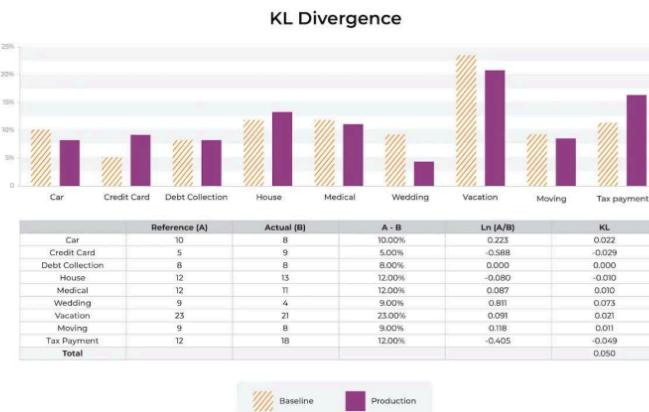
Pendekatan *single-metric* juga rentan terhadap *false sense of stability*, yaitu kondisi di mana satu metrik utama tampak stabil sementara aspek lain dari perilaku model telah mengalami perubahan signifikan. Dalam konteks MLOps, kondisi ini berbahaya karena dapat menyebabkan keterlambatan dalam pengambilan keputusan mitigasi, seperti *rollback* model, sehingga risiko operasional meningkat [10][15].

Tabel 2.3 Perbandingan Single vs Multi-Metric

Aspek	Single-Metric	Multi-Metric
Dimensi	Tunggal	Multidimensi
Sensitivitas	Rendah	Tinggi
Interpretasi	Terbatas	Lebih informatif
Rollback readiness	Lemah	Kuat

2.5 Multi-Metrics Monitoring dan Stability Metrics

Pendekatan *multi-metrics* monitoring muncul sebagai respons terhadap keterbatasan pendekatan *single-metric* dalam merepresentasikan degradasi model yang bersifat multidimensi. Dalam lingkungan produksi yang dinamis, perilaku model pembelajaran mesin tidak hanya berubah pada satu aspek, melainkan pada berbagai dimensi secara simultan, termasuk stabilitas distribusi probabilitas, tingkat kepercayaan prediksi, proporsi kelas, serta karakteristik operasional sistem. Oleh karena itu, monitoring yang efektif harus mampu menangkap perubahan pada berbagai dimensi tersebut secara terintegrasi.



Gambar 2.7 Kerangka Multi-Metrik untuk Monitoring Kesehatan Model

Sumber: <https://arize.com/blog-course/kl-divergence/>

2.5.1 Konsep Monitoring Multidimensional pada Machine Learning

Monitoring multidimensional mengacu pada pendekatan pemantauan model yang melibatkan lebih dari satu metrik untuk merepresentasikan kondisi kesehatan model secara menyeluruh. Pendekatan ini didasarkan pada asumsi bahwa degradasi model merupakan fenomena kompleks yang tidak dapat direduksi menjadi satu indikator tunggal [11][12]. Dalam konteks MLOps, monitoring multidimensional memungkinkan organisasi untuk memperoleh gambaran yang lebih akurat mengenai perilaku model setelah *deployment*. Dengan memantau berbagai aspek secara simultan, sistem dapat mendeteksi perubahan perilaku model yang bersifat halus dan bertahap, yang sering kali luput dari deteksi ketika hanya satu metrik yang digunakan [10][15].

2.5.2 PSI sebagai Indikator Pergeseran Distribusi

PSI merupakan metrik yang digunakan untuk mengukur perbedaan distribusi antara dua populasi data, umumnya antara kondisi *baseline* dan kondisi aktual. PSI awalnya banyak digunakan dalam industri keuangan untuk memantau stabilitas model kredit, namun dalam beberapa tahun terakhir diadopsi secara luas dalam monitoring model pembelajaran mesin [11].

PSI dihitung dengan membandingkan proporsi data pada masing-masing bin distribusi, sehingga mampu mengidentifikasi pergeseran distribusi yang bersifat gradual. Nilai PSI yang tinggi mengindikasikan adanya perubahan distribusi yang signifikan, yang dapat menjadi sinyal awal degradasi model. Dalam konteks penelitian ini, PSI digunakan untuk memantau stabilitas distribusi probabilitas output CNN antar waktu, sehingga memberikan indikasi awal perubahan perilaku model yang tidak selalu tercermin pada metrik performa.

2.5.3 KL Divergence dan Perubahan Distribusi Informasi

KL Divergence merupakan metrik berbasis teori informasi yang mengukur jarak atau perbedaan antara dua distribusi probabilitas. Berbeda dengan PSI yang bersifat bin-based, KL Divergence mengukur kehilangan informasi ketika satu distribusi digunakan untuk merepresentasikan distribusi lain [12]. Dalam konteks monitoring model, KL Divergence sangat sensitif terhadap perubahan distribusi probabilitas output, terutama pada kelas-kelas dengan probabilitas kecil. Sensitivitas ini menjadikan KL Divergence sebagai metrik yang efektif untuk mendeteksi perubahan perilaku model pada tahap awal degradasi. Namun

demikian, karena sifatnya yang sensitif, KL Divergence juga perlu diinterpretasikan secara hati-hati dan dikombinasikan dengan metrik lain agar tidak menghasilkan sinyal palsu (*false alarm*).

2.5.4 *Class Shift* sebagai Indikator Perubahan Pola Prediksi

Class shift terjadi ketika distribusi kelas prediksi model berubah seiring waktu, yang dapat disebabkan oleh perubahan karakteristik data operasional atau munculnya bias baru dalam model [21]. Perubahan proporsi kelas sering kali tidak terdeteksi apabila sistem monitoring hanya berfokus pada metrik agregat seperti akurasi atau *Confidence Ratio*. Oleh karena itu, pemantauan *class shift* menjadi penting untuk memahami bagaimana degradasi model memengaruhi keseimbangan prediksi antar kelas, khususnya pada sistem dengan implikasi risiko yang berbeda untuk masing-masing kelas.

2.5.5 Integrasi Aspek Operasional dalam *Multi-Metrics* Monitoring

Selain aspek performa dan stabilitas distribusi, aspek operasional juga merupakan dimensi penting dalam monitoring model. Latensi inferensi, sebagai contoh, mencerminkan efisiensi sistem dan beban komputasi yang dihadapi model dalam lingkungan produksi [19]. Peningkatan latensi dapat mengindikasikan bahwa model menghadapi input yang semakin kompleks atau tidak sesuai dengan distribusi pelatihan, yang pada akhirnya dapat berdampak pada kualitas layanan.

Integrasi metrik operasional dengan metrik stabilitas dan kepercayaan prediksi memungkinkan analisis degradasi model secara lebih holistik. Pendekatan ini sejalan dengan prinsip monitoring multidimensional yang direkomendasikan dalam praktik MLOps dan standar internasional, yang menekankan pentingnya mempertimbangkan berbagai dimensi risiko dan performa sistem AI secara bersamaan [9][10].

2.5.6 Integrasi *Multi-Metrics* sebagai Dasar *Composite Health Score*

Meskipun masing-masing metrik dalam *multi-metrics* monitoring memberikan informasi yang berharga, interpretasi terpisah terhadap banyak metrik dapat menjadi kompleks dan sulit digunakan sebagai dasar pengambilan keputusan operasional. Oleh karena itu, diperlukan mekanisme integrasi yang mampu menggabungkan berbagai metrik tersebut ke dalam satu indikator yang ringkas namun tetap informatif.

Kebutuhan inilah yang melandasi penggunaan *composite health score* sebagai pendekatan integratif dalam penelitian ini. Dengan menggabungkan metrik stabilitas distribusi (*PSI* dan *KL Divergence*), indikator perubahan pola prediksi (*Class Shift*), metrik kepercayaan (*Confidence Ratio*), serta aspek operasional (*Latency*), *composite health score* diharapkan mampu merepresentasikan kondisi kesehatan model secara holistik dan mendukung pengambilan keputusan yang lebih cepat dan objektif dalam konteks MLOps.

Tabel 2.4 Definisi dan Fungsi Metrik

Metrik	Fungsi	Indikasi
<i>PSI</i>	Perubahan distribusi	<i>Drift</i>
<i>KL Divergence</i>	Divergensi probabilitas	Instabilitas
<i>Confidence Ratio</i>	Stabilitas prediksi	Degradasii
<i>Latency</i>	Kinerja sistem	SLA risk

2.6 *Composite Health Score* dan *Weighted Sum Model* (WSM)

Pendekatan *multi-metrics monitoring* yang dibahas sebelumnya menghasilkan sejumlah indikator yang masing-masing merepresentasikan dimensi tertentu dari kesehatan model. Meskipun pendekatan ini lebih komprehensif dibandingkan *single-metric monitoring*, interpretasi terhadap banyak metrik secara terpisah dapat menjadi kompleks dan kurang efektif dalam mendukung pengambilan keputusan operasional yang cepat. Oleh karena itu, diperlukan suatu mekanisme integratif yang mampu merangkum berbagai indikator tersebut ke dalam satu nilai yang ringkas, informatif, dan mudah diinterpretasikan, tanpa menghilangkan makna dari masing-masing metrik.

2.6.1 Konsep *Composite Health Score* dalam Monitoring Model

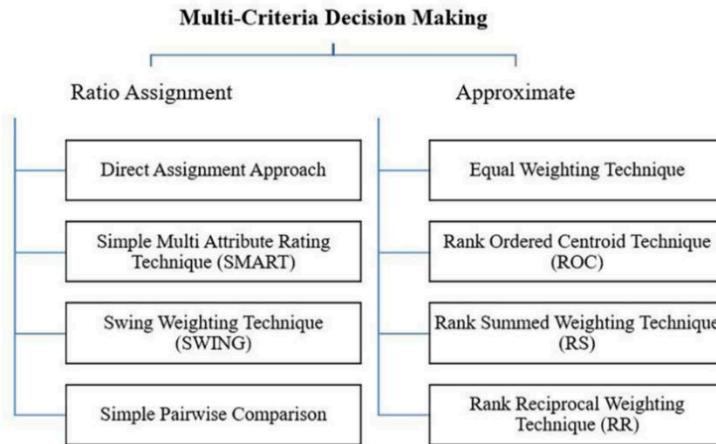
Composite health score merupakan indikator agregat yang dibangun dengan menggabungkan beberapa metrik monitoring untuk merepresentasikan kondisi kesehatan model secara holistik. Konsep ini berangkat dari pemahaman bahwa kesehatan model pembelajaran mesin tidak dapat direduksi menjadi satu dimensi tunggal, melainkan merupakan hasil interaksi dari berbagai aspek, seperti stabilitas

distribusi, tingkat kepercayaan prediksi, perubahan pola kelas, dan kondisi operasional sistem [11][12][15].

Dalam konteks MLOps, *composite health score* berfungsi sebagai *decision-support indicator* yang memungkinkan tim operasional untuk menilai kondisi model secara cepat dan objektif. Alih-alih memantau dan menafsirkan banyak metrik secara simultan, pengambil keputusan dapat menggunakan satu skor komposit sebagai sinyal utama untuk menentukan apakah model masih berada dalam kondisi sehat atau memerlukan tindakan mitigasi seperti rollback.

2.6.2 MCDM dalam Evaluasi Model

Pengembangan *composite health score* secara konseptual sejalan dengan pendekatan MCDM, yaitu kerangka pengambilan keputusan yang melibatkan lebih dari satu kriteria yang saling melengkapi. MCDM banyak digunakan dalam berbagai domain, seperti manajemen risiko, evaluasi kinerja sistem, dan pengambilan keputusan strategis, ketika satu indikator tunggal tidak cukup untuk merepresentasikan kompleksitas permasalahan [13]. Dalam konteks monitoring model pembelajaran mesin, setiap metrik—seperti *PSI*, *KL Divergence*, *Confidence Ratio*, dan *Latency*—dapat dipandang sebagai kriteria yang memiliki tingkat kepentingan berbeda terhadap kesehatan model secara keseluruhan. Oleh karena itu, penggunaan pendekatan MCDM memungkinkan integrasi berbagai metrik tersebut ke dalam satu kerangka evaluasi yang konsisten dan terstruktur.



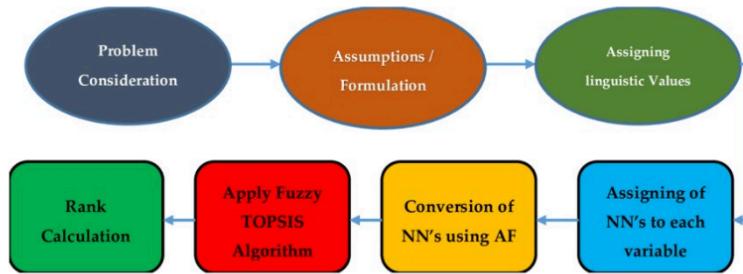
Gambar 10. Taxonomy MCDM

Sumber: https://www.researchgate.net/figure/Taxonomy-of-Multi-Criteria-Decision-Analysis-techniques_fig1_355950271

2.6.3 WSM sebagai Metode Integrasi

WSM merupakan salah satu metode MCDM yang paling sederhana dan banyak digunakan karena kemudahannya dalam implementasi dan interpretasi. Dalam WSM, setiap kriteria dinormalisasi ke dalam skala yang seragam dan kemudian dikalikan dengan bobot tertentu sesuai tingkat kepentingannya. Skor akhir diperoleh dari penjumlahan hasil perkalian antara nilai kriteria dan bobotnya masing-masing [13].

Pemilihan WSM dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, WSM bersifat transparan dan mudah dipahami, sehingga sesuai untuk konteks pengambilan keputusan operasional dalam MLOps. Kedua, WSM memungkinkan fleksibilitas dalam penyesuaian bobot kriteria sesuai kebutuhan dan karakteristik sistem. Ketiga, WSM telah terbukti efektif dalam berbagai aplikasi pengambilan keputusan multiatribut, sehingga memberikan dasar metodologis yang kuat untuk pengembangan *composite health score*.



Gambar 2.9 Tahapan Perhitungan WSM

Sumber: https://www.researchgate.net/figure/Flowchart-for-generalized-fuzzy-TOPSIS-32-Weighted-Sum-Model-WSM-Algorithm-27_fig2_349644160

Tabel 2.5 Skema Bobot WSM

Kriteria	Bobot	Justifikasi
Stability	0.35	Indikator utama drift
Confidence	0.20	Kualitas prediksi
Latency	0.20	Dampak operasional
Error Rate	0.15	Akurasi
Business Impact	0.10	Risiko bisnis

2.6.4 Normalisasi dan Pembobotan Metrik Monitoring

Salah satu tantangan utama dalam pengembangan *composite health score* adalah perbedaan skala dan karakteristik antar metrik. Metrik seperti PSI dan KL *Divergence* meningkat seiring meningkatnya degradasi, sedangkan *Confidence Ratio* menurun ketika degradasi terjadi. Selain itu, metrik operasional seperti *Latency* memiliki satuan dan interpretasi yang berbeda dari metrik probabilistik. Untuk mengatasi perbedaan tersebut, diperlukan proses normalisasi agar seluruh metrik berada pada skala yang sebanding sebelum digabungkan. Normalisasi juga memungkinkan interpretasi skor komposit yang konsisten, di mana nilai yang lebih tinggi atau lebih rendah secara eksplisit merepresentasikan kondisi kesehatan model yang lebih baik atau lebih buruk.

Pembobotan metrik dalam WSM dilakukan untuk merefleksikan tingkat kepentingan relatif masing-masing kriteria terhadap kesehatan model. Dalam konteks monitoring MLOps, metrik stabilitas distribusi dan kepercayaan prediksi umumnya memiliki bobot yang lebih besar karena berkaitan langsung dengan kualitas keputusan model, sementara metrik operasional seperti Latency berfungsi sebagai indikator pendukung [15]. Skema pembobotan ini dirancang agar tetap fleksibel dan dapat disesuaikan dengan kebutuhan sistem atau kebijakan organisasi.

2.6.5 Composite Health Score untuk Pengambilan Keputusan

Composite health score yang dihasilkan melalui pendekatan WSM berfungsi sebagai indikator ringkas yang merepresentasikan kondisi kesehatan model secara menyeluruh. Dalam lingkungan MLOps, skor ini dapat digunakan untuk menetapkan ambang batas (*threshold*) tertentu yang memicu tindakan mitigasi, seperti peringatan dini (*alert*) atau rollback model.

Pendekatan ini memungkinkan pengambilan keputusan yang lebih objektif dan konsisten dibandingkan interpretasi manual terhadap banyak metrik secara terpisah. Penggunaan *composite health score* juga mengurangi ketergantungan pada intuisi atau pengalaman individu dalam menilai kondisi model, sehingga meningkatkan transparansi dan akuntabilitas proses operasional.

Pengembangan *composite health score* berbasis WSM dalam penelitian ini tidak dimaksudkan sebagai sekadar agregasi metrik, melainkan sebagai kerangka evaluasi terstruktur yang menjembatani analisis teknis perilaku model dengan kebutuhan pengambilan keputusan operasional dalam MLOps. Pendekatan ini menjadi fondasi utama untuk analisis empiris pada bab selanjutnya, khususnya dalam membandingkan efektivitas deteksi degradasi antara pendekatan single-metric dan multi-criteria health check.

2.7 Standar Internasional dan AI Governance

Penerapan sistem kecerdasan buatan dalam lingkungan produksi tidak hanya menuntut keandalan teknis, tetapi juga kepatuhan terhadap prinsip tata kelola (*governance*) yang menjamin keamanan, transparansi, dan akuntabilitas. Seiring meningkatnya adopsi AI pada sistem kritis, berbagai organisasi internasional mengembangkan standar untuk memastikan bahwa sistem AI dioperasikan secara bertanggung jawab dan berkelanjutan. Dalam konteks ini, mekanisme monitoring model dan deteksi degradasi menjadi komponen penting dalam penerapan AI governance.

2.7.1 AI Governance dan Risiko Operasional Sistem AI

AI *governance* merujuk pada kerangka kebijakan, proses, dan mekanisme kontrol yang dirancang untuk memastikan bahwa sistem AI beroperasi sesuai dengan tujuan yang diharapkan serta meminimalkan risiko teknis, operasional, dan etis. Salah satu risiko utama dalam pengoperasian sistem AI adalah degradasi model yang tidak terdeteksi, yang dapat menyebabkan sistem menghasilkan keputusan yang tidak akurat atau bias tanpa disadari oleh pengelola sistem [10]. Dalam sistem AI yang beroperasi secara kontinu, risiko degradasi model berkaitan langsung dengan aspek keandalan (*reliability*) dan ketahanan operasional (*operational resilience*). Oleh karena itu, AI *governance* menuntut adanya mekanisme

pemantauan berkelanjutan yang mampu mendeteksi perubahan perilaku model secara dini dan memicu tindakan mitigasi yang sesuai.

2.7.2 Standar ISO/IEC 23053 dan ISO/IEC 5338

ISO/IEC 23053 merupakan standar yang memberikan panduan umum mengenai kerangka kerja sistem AI, termasuk aspek desain, implementasi, dan evaluasi. Standar ini menekankan bahwa sistem AI harus dipantau sepanjang siklus hidupnya, khususnya setelah *deployment*, untuk memastikan bahwa performa dan perilaku sistem tetap berada dalam batas yang dapat diterima [9].

Sementara itu, ISO/IEC 5338 secara lebih spesifik membahas pengelolaan siklus hidup sistem AI, termasuk aspek pemantauan (*monitoring*), pengendalian risiko, dan pemeliharaan sistem. Standar ini menegaskan bahwa evaluasi sistem AI tidak boleh hanya berfokus pada performa awal, tetapi harus mencakup pemantauan berkelanjutan terhadap perubahan kondisi operasional dan data yang dihadapi sistem [10]. Kedua standar tersebut menekankan pentingnya pendekatan evaluasi multidimensional yang mencakup performa, stabilitas, dan aspek operasional. Dengan demikian, penggunaan pendekatan *multi-criteria health check* yang mengintegrasikan berbagai metrik monitoring sejalan dengan rekomendasi yang terkandung dalam standar ISO/IEC tersebut.

2.7.3 Keterkaitan Monitoring Multi-Metrik dengan Prinsip *Governance*

Pendekatan monitoring berbasis multi-metrik memberikan dasar teknis yang kuat untuk penerapan prinsip AI governance. Dengan memantau berbagai dimensi perilaku model, organisasi dapat memperoleh transparansi yang lebih baik mengenai kondisi sistem AI yang dioperasikan. Transparansi ini penting untuk memastikan bahwa keputusan yang dihasilkan oleh sistem AI dapat dipertanggungjawabkan dan diaudit apabila diperlukan.

Selain itu, monitoring multi-metrik mendukung prinsip pencegahan risiko (risk prevention) dalam AI governance. Dengan mendeteksi degradasi model secara dini, organisasi dapat melakukan tindakan mitigasi sebelum dampak negatif terjadi. Pendekatan ini sejalan dengan prinsip kehati-hatian (precautionary principle) yang banyak diadopsi dalam kerangka tata kelola sistem teknologi berisiko tinggi.



Gambar 12. Mapping Monitoring terhadap ISO/IEC

Sumber: <https://gdprlocal.com/a-guide-to-iso-iec-42001-implementation>

2.7.4 Composite Health Score sebagai Implementasi Teknis AI Governance

Dalam konteks penelitian ini, *composite health score* diposisikan sebagai mekanisme teknis yang mendukung penerapan AI governance dalam praktik. Dengan merangkum berbagai metrik monitoring ke dalam satu indikator komposit, sistem dapat menyediakan sinyal yang jelas dan terdokumentasi mengenai kondisi kesehatan model. Indikator ini dapat digunakan sebagai dasar pengambilan keputusan operasional, seperti *rollback*, serta sebagai artefak audit untuk menunjukkan bahwa sistem AI telah dipantau secara berkelanjutan sesuai dengan standar yang berlaku.

Pendekatan ini menjembatani kesenjangan antara rekomendasi normatif dalam standar internasional dan implementasi teknis di lapangan. Dengan demikian, penelitian ini tidak hanya berkontribusi pada aspek teknis monitoring model, tetapi juga pada penerapan prinsip AI *governance* yang lebih kuat dan terstruktur dalam lingkungan MLOps.

Tabel 2.6 Kesesuaian Penelitian dengan Standar

Standar	Prinsip	Relevansi
ISO/IEC 5338	Lifecycle AI	Monitoring

ISO/IEC 23894	Risk AI	Early warning
ISO/IEC 27001	Security	Logging

2.8 Penelitian Terdahulu dan Research Gap

Kajian terhadap penelitian terdahulu dilakukan untuk memahami perkembangan pendekatan dalam monitoring model pembelajaran mesin, khususnya terkait degradasi model, deteksi *drift*, dan mekanisme mitigasi dalam lingkungan produksi. Analisis ini bertujuan untuk mengidentifikasi keterbatasan pendekatan yang telah ada serta menemukan celah penelitian (*research gap*) yang menjadi dasar bagi pengembangan pendekatan yang diusulkan dalam penelitian ini.

2.8.1 Penelitian Terdahulu tentang Degradasi Model dan *Drift Detection*

Berbagai penelitian terdahulu telah membahas fenomena degradasi model dan drift dalam sistem pembelajaran mesin. Studi-studi awal banyak berfokus pada deteksi data *drift* dan *concept drift* melalui analisis perubahan distribusi data input atau performa model [5][6][17]. Pendekatan yang digunakan umumnya melibatkan pengujian statistik terhadap distribusi data atau pemantauan penurunan akurasi sebagai indikator terjadinya degradasi.

Dalam konteks *computer vision*, beberapa penelitian menyoroti sensitivitas CNN terhadap perubahan kualitas visual input, seperti blur, perubahan pencahayaan, dan *noise* [21][27][29]. Penelitian-penelitian tersebut menunjukkan bahwa degradasi visual dapat menyebabkan penurunan performa model, namun sebagian besar evaluasi masih berfokus pada metrik performa tradisional dan dilakukan dalam skenario eksperimental yang terkontrol.

⁴⁴
Tabel 2.7 Penelitian Terdahulu

No	Penulis & Tahun	Judul Penelitian	Fokus Penelitian	Metode / Teknologi	Hasil Utama	Keterbatasan	Relevansi dengan Penelitian Ini
1	Krizhevsky et al., 2012 [1]	<i>ImageNet Classification</i>	Kinerja CNN pada	CNN (AlexNet)	CNN unggul pada akurasi visual	Evaluasi statis (trainin	Dasar penggunaan CNN,

No	Penulis & Tahun	Judul Penelitian	Fokus Penelitian	Metode / Teknologi	Hasil Utama	Keterbatasan	Relevansi dengan Penelitian Ini
		<i>with Deep CNNs</i>	klasifikasi citra			g/testin g)	belum bahas degradasi pasca-deployment
2	Howard et al., 2019 [3]	<i>Searching for MobileNetV3</i>	Efisiensi CNN untuk deployment	MobileNetV3	Model ringan & cepat	Tidak memba has monito ring produksi	Arsitektur model yang digunakan
3	Lu & Xu, 2020 [5]	<i>Understanding Model Decay in Production ML Systems</i>	Model decay di produksi	Analisis performa temporal	Model mengalami penurunan kinerja	Fokus perorma, bukan distribusi output	Landasan degradasi model pasca-deployment
40 4	Gama et al., 2014 [6]	<i>A Survey on Concept Drift Adaptation</i>	Concept drift	Survey adaptasi	Drift umum pada data stream	Fokus adaptasi, bukan monitoring	Pembeda scope (penelitian ini fokus monitoring)
5	Widmer & Kubat, 1996 [16]	<i>Learning in the Presence of Concept Drift</i>	Pembelajaran di bawah drift	Incremental learning	Drift memengaruhi akurasi	Tidak bahas deployment	Referensi konseptual drift

No	Penulis & Tahun	Judul Penelitian	Fokus Penelitian	Metode / Teknologi	Hasil Utama	Keterbatasan	Relevansi dengan Penelitian Ini
						moderan	
6	Žliobaitė, 2010 [17]	<i>Learning under Concept Drift</i>	Definisi & klasifikasi drift	Literature review	Drift bersifat gradual	Tidak bahas MLOps	Landasan klasifikasi drift
7	Barandela et al., 2002 [21]	<i>Quality Drift, Data Drift, and Concept Drift</i>	Drift kualitas data	Analisis konseptual	Kualitas data memicu degradasi	Tanpa mekanisme monitoring	Mendukung fokus degradasi visual
8	Carlini et al., 2019 [27]	<i>Adversarial Examples Are Not Bugs</i>	Sensitivitas model visual	Analisis robustness	CNN sensitif terhadap gangguan	Fokus adversarial	Relevant untuk degradasi kualitas input
9	Goodfellow et al., 2016 [29]	<i>Deep Learning</i>	Fondasi DL	Teori & praktik DL	Model sensitif distribusi data	Bukan monitoring produksi	Landasan teoretis CNN
10	Plumb et al., 2020 [31]	<i>Improving Drift Detection via Probabilistic Inference</i>	Drift berbasis probabilitas	Probabilistic inference	Deteksi drift lebih dini	Tidak operasional (rollback)	Basis monitoring probabilistik

No	Penulis & Tahun	Judul Penelitian	Fokus Penelitian	Metode / Teknologi	Hasil Utama	Keterbatasan	Relevansi dengan Penelitian Ini
11	Kullback & Leibler, 1951 [12]	<i>On Information and Sufficiency</i>	Divergensi distribusi	KL Divergence	Ukur perbedaan distribusi	Bukan khusus ML	Dasar KL Divergence
12	Mitchell, 1997 [11]	<i>Machine Learning</i>	Konsep umum ML	Teori ML	Fondasi evaluasi model	Tidak bahas deploy ment	Landasan konseptual monitoring
13	Sculley et al., 2015 [8]	<i>Hidden Technical Debt in ML Systems</i>	Risiko sistem ML	Analisis sistem	Monitoring krusial	Tidak detail metrik	Landasan MLOps
14	Breck et al., 2017 [15]	<i>The ML Test Score</i>	Production readiness	Checklist MLOps	Monitoring wajib	Tidak implementatif	Mendukung rollback readiness
15	Baier, 2021 [19]	<i>Monitoring ML Models in Production Systems</i>	Monitoring produksi	Survey	Monitoring multidimensi	Tanpa agregasi komposit	Justifikasi multi-metric
16	Bhatt et al., 2021 [14]	<i>Explainable ML in Production</i>	Kepercayaan prediksi	Confidence-based metrics	Confidence sensitif degradasi	Bukan multi-metric	Baseline single-metric
17	Hwang & Yoon, 1981 [13]	<i>Multiple Attribute Decision Making</i>	MCDM	Weighted Sum Model	Agregasi multi-kriteria	Bukan ML	Dasar composite score

No	Penulis & Tahun	Judul Penelitian	Fokus Penelitian	Metode / Teknologi	Hasil Utama	Keterbatasan	Relevansi dengan Penelitian Ini
18	²⁸ ISO/IEC 23053, 2022 [9]	<i>Framework for AI Systems Using ML</i>	Kerangka AI	Standar AI	Evaluasi lifecycle	Normatif	Landasan governance
19	ISO/IEC 5338, 2023 [10]	<i>AI Engineering— Monitoring & Lifecycle</i>	Monitoring AI	Standar internasional	Monitoring multidimensi	Tidak teknis	Validasi tata kelola

2.8.2 Pendekatan Monitoring Berbasis *Single-Metric* dan Keterbatasannya

Sebagian besar sistem monitoring model yang dibahas dalam literatur masih mengandalkan pendekatan *single-metric*, seperti akurasi, F1-score, atau metrik performa lainnya [12][31]. Pendekatan ini relatif mudah diimplementasikan dan memberikan gambaran langsung mengenai performa model terhadap data berlabel.

Namun, sejumlah penelitian mengungkapkan bahwa metrik performa tradisional sering kali gagal mendeteksi degradasi model secara dini karena bersifat sebagai indikator terlambat (*lagging indicator*) [31]. Ketergantungan pada ketersediaan label ground truth membatasi penerapan pendekatan ini dalam lingkungan produksi yang bersifat real-time dan minim label. Meskipun beberapa studi mulai mengeksplorasi metrik berbasis kepercayaan prediksi sebagai indikator alternatif, pendekatan tersebut umumnya masih digunakan secara terpisah dan belum dikaitkan dengan pengambilan keputusan operasional.

2.8.3 Penelitian tentang Monitoring Multi-Metrik dan *Composite Score*

Seiring dengan meningkatnya kompleksitas sistem AI, beberapa penelitian mulai mengusulkan penggunaan stability metrics seperti PSI dan KL Divergence untuk memantau perubahan distribusi data atau output model [11][12]. Pendekatan ini terbukti lebih sensitif terhadap perubahan awal distribusi dibandingkan metrik performa tradisional.

Selain itu, terdapat penelitian yang mengusulkan penggabungan beberapa metrik ke dalam indikator komposit untuk mengevaluasi kondisi sistem secara keseluruhan. Namun, sebagian besar penelitian tersebut masih berfokus pada evaluasi performa atau stabilitas secara umum, tanpa mengaitkannya secara eksplisit dengan konteks MLOps dan pengambilan keputusan operasional seperti rollback model. Penggunaan metode MCDM dalam monitoring model juga masih terbatas dan jarang dievaluasi secara empiris dalam skenario degradasi dunia nyata.

4.2.8.4 Research Gap

Berdasarkan kajian terhadap penelitian terdahulu, dapat diidentifikasi beberapa celah penelitian yang relevan dengan topik penelitian ini. Pertama, sebagian besar penelitian mengenai degradasi model masih berfokus pada deteksi drift atau penurunan performa secara terpisah, tanpa pendekatan integratif yang mampu merepresentasikan kesehatan model secara holistik. Kedua, pendekatan monitoring yang ada masih didominasi oleh penggunaan satu metrik utama, sehingga kurang sensitif terhadap degradasi model yang bersifat multidimensi dan bertahap.

Ketiga, meskipun stability metrics seperti PSI dan KL Divergence telah banyak dibahas dalam literatur, integrasi metrik-metrik tersebut dengan indikator kepercayaan prediksi, perubahan proporsi kelas, dan aspek operasional dalam satu kerangka evaluasi yang terstruktur masih sangat terbatas. Keempat, keterkaitan antara hasil monitoring model dan pengambilan keputusan operasional dalam MLOps, khususnya terkait pemicu rollback model, belum banyak dieksplorasi secara sistematis dalam penelitian terdahulu.

Dengan demikian, terdapat kebutuhan akan pendekatan monitoring model yang tidak hanya mampu mendeteksi degradasi secara dini, tetapi juga relevan secara operasional dan selaras dengan praktik MLOps. Penelitian ini mengisi celah tersebut dengan mengusulkan dan mengevaluasi pendekatan multi-criteria health check berbasis composite health score menggunakan metode WSM, serta membandingkannya dengan pendekatan single-metric *Confidence Ratio* dalam konteks kecepatan deteksi degradasi dan kesiapan rollback model.

2.9 Posisi dan Kontribusi Penelitian

Berdasarkan kajian pustaka dan analisis research gap yang telah dibahas sebelumnya, posisi penelitian ini dapat ditempatkan pada irisan antara kajian degradasi model pembelajaran mesin, praktik MLOps, dan kebutuhan pengambilan keputusan operasional dalam sistem AI yang beroperasi secara kontinu. Penelitian ini tidak berfokus pada pengembangan arsitektur model baru atau peningkatan performa klasifikasi semata, melainkan pada pemahaman dan pengelolaan perilaku model setelah deployment dalam lingkungan produksi yang dinamis.

Secara konseptual, penelitian ini memposisikan degradasi model sebagai fenomena multidimensi yang tidak dapat direpresentasikan secara memadai oleh satu indikator tunggal. Dengan demikian, penelitian ini mengambil posisi kritis terhadap pendekatan single-metric monitoring yang masih dominan dalam praktik, dan mengusulkan pendekatan multi-criteria health check sebagai kerangka evaluasi yang lebih sesuai untuk kebutuhan operasional MLOps. Dalam kerangka ini, kesehatan model dipahami sebagai hasil interaksi antara stabilitas distribusi probabilitas, tingkat kepercayaan prediksi, perubahan pola kelas, dan kondisi operasional sistem.

Dari sisi metodologis, penelitian ini berkontribusi dengan mengadopsi dan menerapkan pendekatan MCDM, khususnya WSM, sebagai mekanisme integrasi berbagai metrik monitoring ke dalam satu *composite health score*. Berbeda dengan penelitian terdahulu yang umumnya membahas metrik monitoring secara terpisah, penelitian ini menyajikan kerangka evaluasi terstruktur yang memungkinkan berbagai metrik dengan karakteristik berbeda digabungkan secara sistematis dan transparan. Pendekatan ini memberikan kontribusi metodologis dalam konteks monitoring model pembelajaran mesin yang masih relatif jarang dieksplorasi.

Penelitian ini berkontribusi dengan melakukan evaluasi komparatif antara pendekatan *single-metric* monitoring berbasis *Confidence Ratio* dan pendekatan *multi-criteria health check* berbasis *composite health score* dalam skenario degradasi data yang disimulasikan secara terkontrol. Evaluasi difokuskan pada kemampuan masing-masing pendekatan dalam mendeteksi degradasi model secara dini, bukan semata-mata pada penurunan performa akhir..

Penelitian ini memberikan kontribusi praktis dengan mengaitkan hasil monitoring model secara langsung dengan pengambilan keputusan operasional dalam pipeline MLOps, khususnya terkait kesiapan dan efektivitas proses *rollback* model. Dengan memposisikan composite health score sebagai indikator objektif untuk memicu rollback, penelitian ini menjembatani kesenjangan antara analisis teknis perilaku model dan kebutuhan pengelolaan sistem AI yang andal dan berkelanjutan di lingkungan produksi.

Secara keseluruhan, kontribusi utama penelitian ini terletak pada pengembangan kerangka monitoring model yang lebih holistik, operasional, dan selaras dengan prinsip AI *governance*. Dengan mengintegrasikan aspek teknis, metodologis, dan operasional, penelitian ini diharapkan dapat menjadi referensi bagi pengembangan sistem MLOps yang lebih adaptif dalam menghadapi degradasi model, sekaligus memberikan dasar yang kuat bagi penelitian lanjut dalam bidang monitoring dan pengelolaan sistem AI.

Tabel 2.8 Posisi Penelitian terhadap *State-of-the-Art*

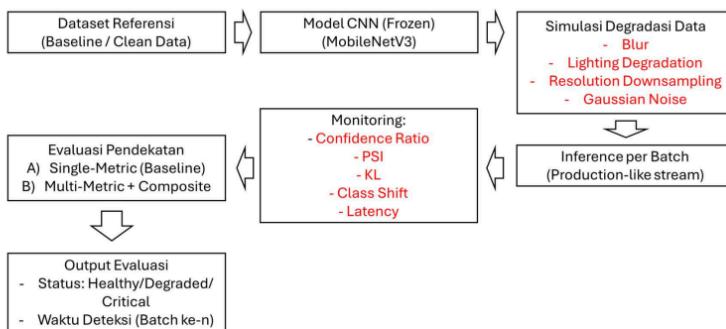
No	State-of-the-Art / Literatur	Cakupan & Pendekatan Utama	Keterbatasan Penelitian Terdahulu	Posisi dan Kontribusi Penelitian Ini
1	Studi CNN Klasik [1][3]	Evaluasi performa statis CNN	Tidak bahas degradasi produksi	Fokus monitoring pasca-deployment
2	Analisis Model Decay [5]	Penurunan performa temporal	Metrik performa (lagging)	Deteksi dini berbasis probabilistik
3	Survey Concept Drift [6][17]	Deteksi & adaptasi drift	Fokus adaptasi	Monitoring tanpa retraining
4	Quality/Data Drift [21]	Dampak kualitas data	Tanpa indikator operasional	Health indicator berbasis data
5	Robustness CNN [27][29]	Sensitivitas input	Bukan monitoring	Dasar skenario degradasi

6	Single-Metric Monitoring [14]	Confidence-based	Tidak holistik	Dibandingkan sebagai baseline
7	Monitoring Produksi ML [19]	Multi-metric terpisah	Tanpa agregasi	Composite health score
8	Best Practice MLOps [8][15]	Reliability & readiness	Normatif	Implementasi teknis konkret
9	MCDM / WSM [13]	Agregasi multi-kriteria	Bukan domain ML	Adaptasi untuk health score
10	Evaluasi ML Umum [11]	Teori evaluasi	Tidak temporal	Evaluasi batch-based
11	Monitoring Berbasis Performa	Akurasi/F1	Lagging & label-dependent	Monitoring tanpa label
12	Monitoring Terpisah	Banyak metrik	Sulit keputusan	Indikator komposit tunggal
13	Rollback Ad-hoc	Keputusan manual	Tidak objektif	Rollback berbasis threshold

12
BAB III
METODOLOGI PENELITIAN

3.1 Pendekatan dan Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimental kuantitatif, yang bertujuan untuk menganalisis perubahan perilaku model klasifikasi berbasis CNN ketika dihadapkan pada kondisi degradasi data yang disimulasikan secara terkontrol. Pendekatan eksperimental dipilih karena memungkinkan peneliti mengamati hubungan kausal antara variabel degradasi data dan perubahan metrik monitoring model secara terukur dan objektif, sebagaimana direkomendasikan dalam studi-studi evaluasi sistem pembelajaran mesin pasca-deployment [6][31].



Gambar 13. Desain Penelitian Eksperimental untuk Deteksi Degradasi Model
Sumber: Penulis, 2025

Desain penelitian ini mengadopsi eksperimen terkontrol (*controlled experiment*), di mana satu arsitektur model dan dataset yang sama digunakan pada berbagai skenario degradasi data. Pendekatan ini bertujuan untuk meminimalkan variabel pengganggu sehingga setiap perubahan pada metrik monitoring dapat diatribusikan secara langsung pada faktor degradasi data, bukan pada perbedaan model atau konfigurasi pelatihan [5]. Desain eksperimen semacam ini umum digunakan dalam penelitian degradasi model dan drift analysis karena mampu memberikan interpretasi yang lebih jelas terhadap penyebab perubahan perilaku model [17].

Dalam konteks penelitian ini, degradasi model diposisikan sebagai fenomena pasca-deployment yang bersifat gradual dan multidimensi. Oleh karena itu, desain penelitian tidak diarahkan pada peningkatan performa model melalui pelatihan ulang atau optimasi arsitektur, melainkan pada evaluasi mekanisme monitoring yang mampu mendeteksi perubahan perilaku model secara dini. Pendekatan ini sejalan dengan praktik MLOps yang menekankan pentingnya pemantauan berkelanjutan dan mitigasi risiko operasional sebelum dampak degradasi menjadi signifikan [8][10][15].

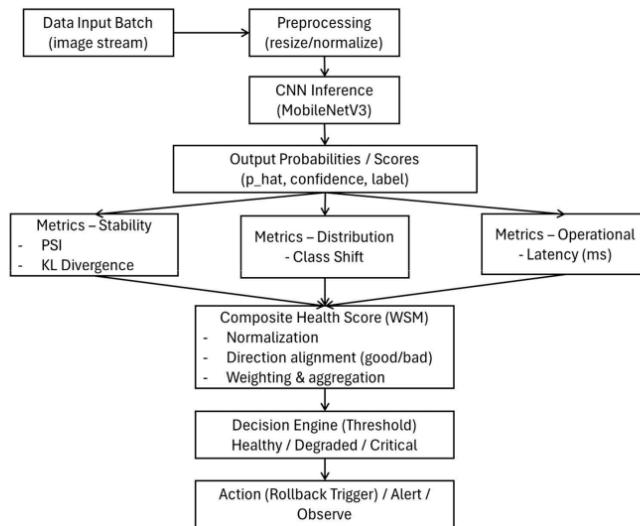
Eksperimen dilakukan dengan mensimulasikan aliran data operasional dalam bentuk pemrosesan berbasis batch, di mana setiap *batch* merepresentasikan periode operasional tertentu. Pendekatan *batch-level monitoring* dipilih karena lebih realistik untuk merepresentasikan sistem produksi dibandingkan evaluasi statis satu kali, serta telah banyak digunakan dalam studi *monitoring* model untuk mengamati perubahan perilaku model secara temporal [31]. Pendekatan ini memungkinkan identifikasi pola degradasi yang bersifat bertahap dan tidak langsung terlihat melalui evaluasi performa statis.

Untuk menjawab rumusan masalah penelitian, desain eksperimen dibangun dengan dua jalur evaluasi utama. Jalur pertama menggunakan *Confidence Ratio* sebagai representasi pendekatan single-metric monitoring, yang berfungsi sebagai baseline dan merepresentasikan praktik monitoring sederhana yang umum digunakan [12][14]. Jalur kedua menggunakan pendekatan *multi-criteria health check*, yang mengintegrasikan beberapa metrik monitoring—meliputi PSI, KL Divergence, *class shift*, *Confidence Ratio*, dan *latency*—ke dalam satu *composite health score* menggunakan metode WSM, sebagaimana direkomendasikan dalam kerangka *multi-criteria decision making* [11][13].

Perbandingan antara kedua pendekatan tersebut difokuskan pada kecepatan dan konsistensi deteksi degradasi model, bukan semata-mata pada nilai performa akhir. Fokus ini sejalan dengan tujuan operasional MLOps, di mana deteksi dini degradasi memiliki peran penting dalam mendukung pengambilan keputusan mitigasi seperti rollback model [15]. Dengan demikian, desain penelitian ini secara eksplisit menghubungkan mekanisme monitoring model dengan kebutuhan pengelolaan sistem AI yang andal dan berkelanjutan.

3.2 Arsitektur Sistem Monitoring Model

Arsitektur sistem monitoring model dalam penelitian ini dirancang untuk merepresentasikan alur operasional model pembelajaran mesin di lingkungan produksi. Perancangan arsitektur tidak hanya mempertimbangkan aspek inferensi model, tetapi juga mencakup proses pengumpulan metrik, evaluasi kesehatan model, serta pengambilan keputusan operasional berbasis hasil monitoring. Pendekatan ini sejalan dengan praktik MLOps yang menekankan integrasi antara model inference, monitoring berkelanjutan, dan mekanisme mitigasi risiko [8][9][10].



Gambar 3.2 Arsitektur Sistem Monitoring Model dalam Lingkungan MLOps
Sumber: Penulis, 2025

Secara konseptual, arsitektur sistem monitoring model pada penelitian ini terdiri dari beberapa lapisan utama, yaitu lapisan input data, lapisan inferensi model, lapisan ekstraksi metrik, lapisan evaluasi kesehatan model, dan lapisan pengambilan keputusan operasional. Pembagian lapisan ini bertujuan untuk memisahkan tanggung jawab masing-masing komponen, sehingga sistem lebih mudah dianalisis, dikembangkan, dan direproduksi.

Tabel 3.1 Komponen Arsitektur Monitoring

Komponen	Fungsi
Data Input Layer	Menerima data operasional
Inference Layer	Prediksi probabilistik
Metrics Layer	Hitung metrik monitoring
Health Evaluation	Hitung composite score
Decision Layer	Trigger rollback

3.2.1 Lapisan Input Data dan Simulasi Aliran Produksi

Lapisan input data berfungsi sebagai sumber data yang merepresentasikan aliran data operasional yang diterima model setelah deployment. Dalam penelitian ini, data input disusun dalam bentuk *batch* yang merepresentasikan periode waktu tertentu, sehingga memungkinkan analisis perubahan perilaku model secara temporal. Pendekatan batch-based streaming ini umum digunakan dalam penelitian monitoring model karena mampu menyeimbangkan antara realisme sistem produksi dan kemudahan analisis eksperimental [31]. Pada lapisan ini, berbagai skenario degradasi data diterapkan secara terkontrol untuk mensimulasikan kondisi lingkungan operasional yang dinamis. Dengan memisahkan lapisan input data dari lapisan lainnya, penelitian ini memastikan bahwa efek degradasi dapat diamati secara langsung pada perubahan metrik monitoring tanpa dipengaruhi oleh faktor lain.

3.2.2 Lapisan Inferensi Model

Lapisan inferensi model merupakan komponen inti yang menjalankan proses prediksi menggunakan model CNN yang telah dilatih sebelumnya. Model menerima data input dari lapisan sebelumnya dan menghasilkan output berupa distribusi probabilitas kelas melalui fungsi *softmax*. Output probabilistik ini menjadi dasar bagi perhitungan berbagai metrik monitoring, khususnya metrik berbasis kepercayaan dan stabilitas distribusi [12]. Dalam konteks MLOps, pemisahan lapisan inferensi dari lapisan monitoring memungkinkan sistem untuk memantau perilaku model tanpa mengganggu proses prediksi utama. Pendekatan ini sejalan dengan prinsip non-intrusive monitoring, di mana sistem monitoring berjalan secara paralel dengan sistem inferensi produksi [10].

3.2.3 Lapisan Ekstraksi dan Perhitungan Metrik Monitoring

Lapisan ekstraksi metrik bertanggung jawab untuk menghitung berbagai metrik monitoring berdasarkan output model dan karakteristik operasional sistem. Metrik yang dihitung pada lapisan ini mencakup *Confidence Ratio*, PSI, KL *Divergence*, *class shift*, serta *latency* inferensi. Perhitungan metrik dilakukan secara terpisah untuk setiap batch data, sehingga memungkinkan observasi perubahan metrik secara bertahap. Pendekatan ini selaras dengan rekomendasi literatur yang menyatakan bahwa degradasi model sering kali muncul secara gradual dan lebih mudah dideteksi melalui pemantauan temporal dibandingkan evaluasi statis [11][31].

3.2.4 Lapisan Evaluasi Kesehatan Model

Lapisan evaluasi kesehatan model berfungsi untuk mengintegrasikan hasil perhitungan metrik monitoring ke dalam satu indikator komposit, yaitu *composite health score*. Pada lapisan ini, nilai metrik yang telah dinormalisasi digabungkan menggunakan metode WSM sesuai dengan bobot yang telah ditentukan [13]. Lapisan ini merepresentasikan *decision-support layer* dalam arsitektur MLOps, di mana informasi teknis dari berbagai metrik diterjemahkan menjadi indikator yang lebih mudah diinterpretasikan. Dengan adanya lapisan evaluasi kesehatan, sistem monitoring dapat memberikan sinyal kondisi model secara ringkas dan konsisten, yang penting untuk mendukung pengambilan keputusan operasional [15].

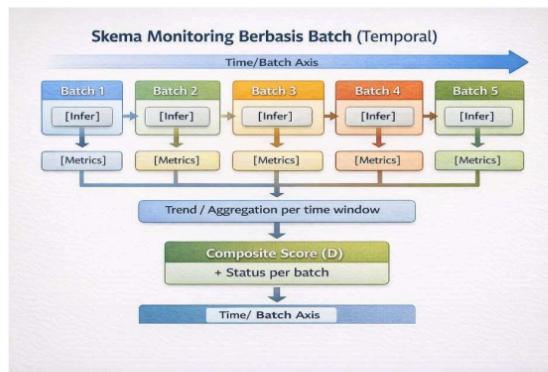
3.2.5 Lapisan Pengambilan Keputusan Operasional

Lapisan pengambilan keputusan operasional merupakan komponen akhir dalam arsitektur sistem monitoring model. Pada lapisan ini, *composite health score* dibandingkan dengan ambang batas (*threshold*) tertentu untuk menentukan apakah model berada dalam kondisi sehat atau memerlukan tindakan mitigasi, seperti rollback. Penggunaan ambang batas berbasis indikator komposit memungkinkan pengambilan keputusan yang lebih objektif dibandingkan interpretasi manual terhadap banyak metrik secara terpisah. Pendekatan ini sejalan dengan praktik MLOps yang menekankan otomatisasi dan konsistensi dalam pengelolaan sistem AI [10][15].

3.2.6 Keterkaitan Arsitektur dengan Tujuan Penelitian

Arsitektur sistem monitoring model yang dirancang dalam penelitian ini secara eksplisit mengaitkan proses inferensi model dengan mekanisme monitoring dan pengambilan keputusan operasional. Dengan struktur berlapis, arsitektur ini memungkinkan analisis yang jelas terhadap hubungan antara degradasi data, perubahan metrik monitoring, dan implikasinya terhadap kesehatan model.

Selain itu, arsitektur ini dirancang agar bersifat modular dan model-agnostic, sehingga pendekatan monitoring yang diusulkan dapat diperluas ke arsitektur model lain atau skenario produksi yang berbeda. Dengan demikian, arsitektur sistem monitoring model ini tidak hanya mendukung kebutuhan eksperimen penelitian, tetapi juga merepresentasikan praktik MLOps yang realistik dan relevan secara operasional.



Gambar 3.3 Skema Monitoring Model Berbasis *Batch*
Sumber: Penulis, 2025

3.3 Model dan Dataset Penelitian

Pemilihan model dan dataset dalam penelitian ini dilakukan dengan mempertimbangkan relevansi terhadap tujuan penelitian, yaitu menganalisis degradasi model dan efektivitas mekanisme monitoring pasca-deployment dalam konteks MLOps. Oleh karena itu, aspek efisiensi komputasi, stabilitas inferensi, serta kesesuaian dengan skenario produksi menjadi pertimbangan utama, bukan semata-mata pencapaian performa klasifikasi tertinggi.

Tabel 3.2 Spesifikasi Model

Aspek	Deskripsi
Arsitektur	MobileNetV3
Tipe	CNN
Output	Binary classification
Status	Frozen model

3.3.1 Arsitektur Model Klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini adalah CNN dengan arsitektur MobileNetV3. MobileNetV3 merupakan arsitektur CNN ringan yang dirancang untuk efisiensi komputasi dan kecepatan inferensi, sehingga banyak digunakan pada sistem dengan keterbatasan sumber daya dan kebutuhan real-time inference [3].

Pemilihan MobileNetV3 didasarkan pada beberapa pertimbangan. Pertama, arsitektur ini menggabungkan *depthwise separable convolution* dan *inverted residual blocks* yang memungkinkan pengurangan kompleksitas komputasi tanpa mengorbankan kemampuan ekstraksi fitur secara signifikan. Kedua, MobileNetV3 dirancang dengan mempertimbangkan *hardware-aware optimization*, sehingga relevan untuk lingkungan produksi yang menuntut efisiensi dan stabilitas inferensi [3].

Dalam konteks penelitian ini, MobileNetV3 tidak dipilih untuk menunjukkan keunggulan performa dibandingkan arsitektur CNN lain, melainkan sebagai representasi model produksi yang realistik. Dengan menggunakan arsitektur yang umum digunakan dalam praktik, hasil analisis degradasi dan monitoring yang diperoleh diharapkan lebih relevan dan mudah digeneralisasikan ke sistem nyata. Selain itu, karakteristik MobileNetV3 yang menghasilkan output probabilistik melalui fungsi softmax menjadikannya sesuai untuk evaluasi metrik berbasis kepercayaan dan stabilitas distribusi [12].

Model dilatih menggunakan data pelatihan yang telah dipersiapkan sebelumnya dan kemudian dipertahankan dalam kondisi tetap (*frozen model*) selama seluruh rangkaian eksperimen degradasi. Pendekatan ini bertujuan untuk memastikan bahwa perubahan metrik monitoring yang diamati benar-benar

disebabkan oleh degradasi data input, bukan oleh perubahan parameter atau adaptasi model.

3.3.2 Dataset dan Definisi Kelas

Dataset yang digunakan dalam penelitian ini terdiri dari data citra yang disusun untuk tugas klasifikasi dua kelas (*binary classification*). Pemilihan skema dua kelas dilakukan untuk menyederhanakan analisis perubahan distribusi probabilitas output dan mempermudah interpretasi metrik stabilitas seperti PSI dan KL Divergence, sebagaimana direkomendasikan dalam penelitian monitoring model berbasis distribusi [11][12]. Dataset dibagi ke dalam beberapa subset, yaitu data pelatihan (*training set*), data validasi (*validation set*), dan data pengujian (*test set*). Data pelatihan dan validasi digunakan untuk membangun dan mengevaluasi performa awal model sebelum deployment, sementara data pengujian digunakan sebagai baseline evaluasi awal. Setelah tahap ini, dataset tambahan digunakan untuk mensimulasikan aliran data operasional yang menjadi objek utama monitoring degradasi model.

Dalam simulasi lingkungan produksi, data disajikan dalam bentuk batch yang merepresentasikan periode operasional tertentu. Setiap batch kemudian dikenai skenario degradasi data yang berbeda untuk mensimulasikan kondisi dunia nyata, seperti perubahan pencahayaan, blur, resolusi rendah, dan noise. Pendekatan ini memungkinkan analisis temporal terhadap perubahan metrik monitoring dan mencerminkan kondisi operasional sistem computer vision yang menerima data secara kontinu [21][27]. Definisi kelas dalam dataset ditetapkan secara konsisten sepanjang eksperimen untuk memastikan bahwa perubahan distribusi prediksi yang diamati benar-benar mencerminkan degradasi perilaku model. Pendekatan ini juga mendukung analisis class shift sebagai salah satu indikator degradasi model, di mana perubahan proporsi prediksi antar kelas dapat diamati secara kuantitatif [21].

Tabel 3.3 Ringkasan Dataset

Dataset	Total Data	Peran
Training	612 files	Pelatihan model
Validation	612 files	Validasi awal
Test	76 files	Baseline performa
Production-like	76 files	Monitoring degradasi

3.3.3 Keterkaitan Model dan Dataset dengan Tujuan Penelitian

Pemilihan MobileNetV3 sebagai model dan penggunaan dataset dua kelas dalam penelitian ini secara langsung mendukung tujuan penelitian, yaitu mengevaluasi efektivitas mekanisme monitoring degradasi model pasca-deployment. Dengan menggunakan model yang efisien dan dataset yang terkontrol, penelitian ini dapat memfokuskan analisis pada perubahan perilaku model, bukan pada kompleksitas arsitektur atau perbedaan skala dataset. Pendekatan ini memungkinkan evaluasi yang lebih jelas terhadap perbedaan antara *single-metric* monitoring dan *multi-criteria health check* dalam mendeteksi degradasi model secara dini. Dengan demikian, model dan dataset yang digunakan tidak hanya berfungsi sebagai sarana eksperimen, tetapi juga sebagai komponen metodologis yang selaras dengan kerangka MLOps dan tujuan operasional penelitian ini.

3.4 Skenario Degradasi Data

Skenario degradasi data dalam penelitian ini dirancang untuk merepresentasikan kondisi nyata yang sering dihadapi oleh sistem *computer vision* setelah *deployment* di lingkungan produksi. Tujuan utama perancangan skenario degradasi bukan untuk merusak performa model secara ekstrem, melainkan untuk mensimulasikan perubahan kualitas data yang bersifat gradual dan realistik, sebagaimana yang umum terjadi pada sistem berbasis kamera di dunia nyata [21][27].

Dalam konteks degradasi model, perubahan kualitas data input merupakan salah satu penyebab utama terjadinya pergeseran distribusi fitur dan output probabilistik model. Oleh karena itu, skenario degradasi dalam penelitian ini difokuskan pada aspek kualitas visual citra, yang secara langsung memengaruhi kemampuan CNN dalam mengekstraksi fitur visual secara konsisten.

Tabel 3.4 Daftar Skenario Degradasi

Skenario	Parameter	Tujuan
Blur	Kernel size	Simulasi getaran
Lighting	Brightness factor	Kondisi pencahayaan
Resolution	Downsampling	Keterbatasan bandwidth

Noise	Gaussian noise	Gangguan sensor
-------	----------------	-----------------

3.4.1 Prinsip Perancangan Skenario Degradasi

Perancangan skenario degradasi data didasarkan pada tiga prinsip utama. Pertama, degradasi harus merepresentasikan kondisi operasional yang realistik, seperti variasi pencahayaan atau penurunan kualitas sensor. Kedua, degradasi diterapkan secara terkontrol dan bertahap, sehingga memungkinkan observasi perubahan metrik monitoring dari waktu ke waktu. Ketiga, degradasi harus tidak mengubah label ground truth, sehingga perubahan perilaku model dapat diatribusikan secara langsung pada kualitas data input, bukan pada perubahan semantik objek [27][29]. Skenario degradasi dalam penelitian ini dirancang untuk menimbulkan perubahan perilaku model yang subtil namun konsisten, sehingga sesuai untuk mengevaluasi efektivitas pendekatan monitoring dalam mendeteksi degradasi secara dini.

3.4.2 Degradasi Blur

Degradasi blur disimulasikan untuk merepresentasikan kondisi citra yang mengalami kehilangan ketajaman akibat pergerakan kamera, getaran perangkat, atau fokus lensa yang tidak optimal. Blur merupakan salah satu bentuk degradasi visual yang paling umum pada sistem computer vision berbasis kamera, terutama pada lingkungan luar ruang atau sistem bergerak [21]. Penerapan blur menyebabkan hilangnya detail spasial pada citra, yang berdampak langsung pada lapisan awal CNN yang bertugas mengekstraksi fitur tepi dan tekstur. Penelitian terdahulu menunjukkan bahwa CNN sangat sensitif terhadap blur, dan perubahan kecil pada tingkat blur dapat memicu perubahan distribusi probabilitas output sebelum penurunan akurasi terlihat secara signifikan [27][29].

3.4.3 Degradasi Penurunan Pencahayaan

Degradasi penurunan pencahayaan (*brightness reduction*) disimulasikan untuk merepresentasikan variasi kondisi pencahayaan yang umum terjadi dalam sistem *computer vision*, seperti perbedaan waktu siang dan malam, bayangan, atau kondisi cuaca tertentu. Variasi pencahayaan merupakan faktor eksternal yang sulit dikontrol dan sering kali tidak sepenuhnya terwakili dalam data pelatihan [21].

Penurunan pencahayaan memengaruhi distribusi intensitas piksel dan kontras citra, yang dapat menyebabkan pergeseran distribusi fitur yang diekstraksi oleh CNN. Studi sebelumnya menunjukkan bahwa perubahan pencahayaan dapat menyebabkan penurunan tingkat kepercayaan prediksi model meskipun prediksi kelas akhir masih benar [27]. Oleh karena itu, skenario ini relevan untuk mengevaluasi sensitivitas metrik berbasis kepercayaan dan stabilitas distribusi.

3.4.4 Degradasi Resolusi Rendah

Degradasi resolusi rendah (*low resolution*) disimulasikan untuk merepresentasikan kondisi penurunan kualitas citra akibat keterbatasan *bandwidth*, kompresi data, atau penggunaan perangkat kamera dengan resolusi rendah. Dalam sistem produksi, resolusi citra sering kali disesuaikan untuk mengoptimalkan penggunaan sumber daya, yang dapat berdampak pada kualitas input model [27]. Penurunan resolusi menyebabkan hilangnya detail visual dan perubahan skala fitur, yang dapat mengganggu konsistensi representasi fitur dalam CNN. Degradasi ini sering kali bersifat gradual dan sulit dideteksi menggunakan metrik performa tradisional, sehingga menjadi skenario yang tepat untuk menguji efektivitas pendekatan monitoring multi-metrik dalam mendeteksi degradasi model secara dini.

3.4.5 Degradasi Noise

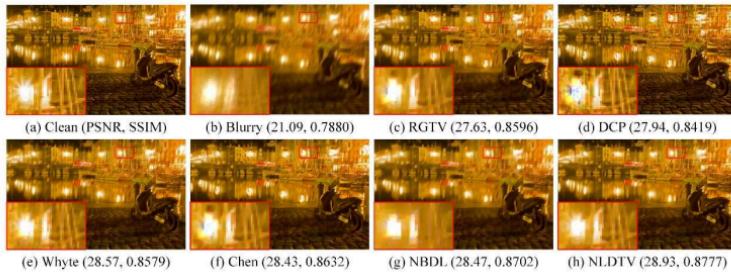
Degradasi noise disimulasikan untuk merepresentasikan gangguan sensorik yang umum terjadi pada sistem kamera, seperti sensor noise, interferensi sinyal, atau kondisi pencahayaan ekstrem. Noise dapat menyebabkan fluktuasi acak pada nilai piksel, yang berdampak pada ketidakstabilan fitur yang diekstraksi oleh CNN [29]. Penambahan noise pada citra input dapat menyebabkan ketidakpastian dalam prediksi model dan perubahan distribusi probabilitas output. Meskipun model masih dapat menghasilkan prediksi yang benar, tingkat kepercayaan dan stabilitas distribusi sering kali menurun. Skenario noise digunakan untuk mengevaluasi kemampuan metrik monitoring dalam menangkap perubahan perilaku model yang bersifat acak namun konsisten secara statistik.

3.4.6 Relevansi Skenario Degradasi terhadap Tujuan Penelitian

Keempat skenario degradasi yang digunakan dalam penelitian ini—blur, penurunan pencahayaan, resolusi rendah, dan noise—dipilih karena

merepresentasikan kondisi umum yang sering dihadapi oleh sistem computer vision di lingkungan produksi. Dengan menerapkan degradasi secara terkontrol dan bertahap, penelitian ini memungkinkan analisis yang jelas terhadap hubungan antara degradasi data input dan perubahan metrik monitoring model.

Skenario degradasi ini juga dirancang untuk mendukung evaluasi perbandingan antara pendekatan single-metric monitoring dan multi-criteria health check. Dengan memunculkan degradasi yang tidak selalu langsung menurunkan performa akhir, skenario ini memberikan konteks yang tepat untuk menilai kemampuan masing-masing pendekatan dalam mendeteksi degradasi model secara dini dan mendukung kesiapan pengambilan keputusan rollback dalam pipeline MLOps.



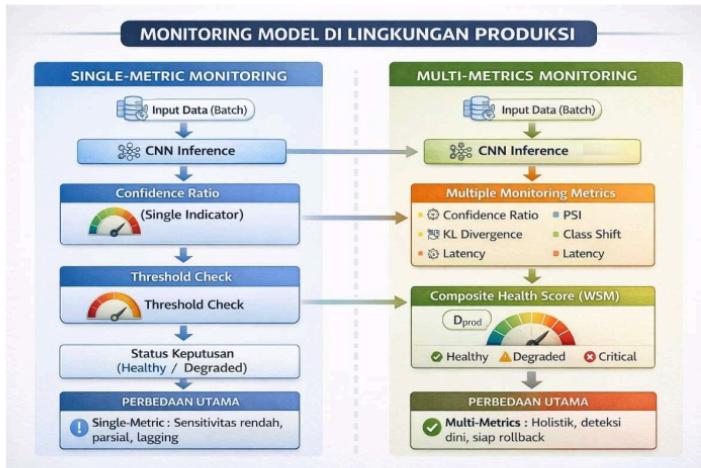
Gambar 3.4 Contoh Skenario Degradasi Data Visual
Sumber: <https://www.mdpi.com/1424-8220/23/8/3784>

3.5 Mekanisme Monitoring dan Metrik Evaluasi

Mekanisme monitoring dalam penelitian ini dirancang untuk mengevaluasi perubahan perilaku model klasifikasi secara sistematis ketika dihadapkan pada kondisi degradasi data. Monitoring tidak hanya difokuskan pada hasil akhir prediksi, tetapi juga pada karakteristik probabilistik dan operasional model yang mencerminkan kondisi kesehatannya secara lebih menyeluruh. Pendekatan ini sejalan dengan praktik Machine Learning Operations (MLOps) yang menekankan pemantauan berkelanjutan terhadap berbagai dimensi perilaku model setelah deployment [8][10][15].

Untuk menjawab tujuan penelitian, mekanisme monitoring dibagi ke dalam dua pendekatan utama, yaitu single-metric monitoring dan multi-metrics

monitoring. Kedua pendekatan ini diterapkan secara paralel pada skenario degradasi yang sama untuk memungkinkan perbandingan yang adil dan terukur.



Gambar 3.5 Single vs Multi-Metrics Monitoring

Sumber: Penulis, 2025

3.5.1 Single-Metric Monitoring: *Confidence Ratio*

Pendekatan *single-metric* monitoring dalam penelitian ini menggunakan *Confidence Ratio* sebagai metrik utama untuk merepresentasikan tingkat kepercayaan model terhadap prediksi yang dihasilkan. *Confidence Ratio* dihitung berdasarkan distribusi probabilitas output model, yang diperoleh melalui fungsi softmax pada lapisan output CNN. Secara konseptual, *Confidence Ratio* merefleksikan seberapa yakin model dalam memilih kelas prediksi dibandingkan alternatif kelas lainnya [12][14].

Pemilihan *Confidence Ratio* sebagai *baseline* didasarkan pada beberapa pertimbangan metodologis. Pertama, *Confidence Ratio* tidak bergantung pada ketersediaan label ground truth, sehingga lebih relevan untuk monitoring model dalam lingkungan produksi yang minim label. Kedua, *Confidence Ratio* bersifat kontinu dan sensitif terhadap perubahan probabilistik, sehingga berpotensi mendeteksi degradasi model lebih awal dibandingkan metrik performa diskrit seperti akurasi [31].

Dalam implementasinya, *Confidence Ratio* dihitung untuk setiap batch data dan kemudian dianalisis secara temporal untuk mengamati tren perubahan tingkat kepercayaan model. Penurunan *Confidence Ratio* secara konsisten diinterpretasikan sebagai indikasi awal degradasi model. Namun demikian, karena *Confidence Ratio* hanya merepresentasikan satu dimensi kesehatan model, pendekatan ini digunakan sebagai baseline banding, bukan sebagai solusi utama dalam penelitian ini.

3.5.2 Multi-Metrics Monitoring

Pendekatan *multi-metrics* monitoring dalam penelitian ini dirancang untuk menangkap degradasi model yang bersifat multidimensi dengan memantau beberapa indikator secara simultan. Metrik yang digunakan dipilih untuk merepresentasikan aspek stabilitas distribusi, perubahan pola prediksi, tingkat kepercayaan model, dan kondisi operasional sistem. Metrik pertama yang digunakan adalah PSI, yang berfungsi untuk mengukur pergeseran distribusi probabilitas output model antara kondisi baseline dan kondisi operasional. PSI digunakan karena kemampuannya dalam mendeteksi perubahan distribusi yang bersifat gradual dan sering digunakan dalam monitoring sistem prediktif [11].

Metrik kedua adalah KL *Divergence*, yang mengukur perbedaan informasi antara dua distribusi probabilitas output. KL *Divergence* bersifat sensitif terhadap perubahan kecil pada distribusi probabilitas dan memberikan indikasi awal perubahan perilaku model yang tidak selalu tercermin pada metrik performa [12]. Metrik ketiga adalah *Class Shift*, yang digunakan untuk mengukur perubahan proporsi prediksi antar kelas dari waktu ke waktu. Perubahan proporsi kelas dapat mengindikasikan adanya bias baru atau pergeseran karakteristik data operasional, meskipun nilai *Confidence Ratio* atau akurasi belum menunjukkan perubahan signifikan [21].

Selain metrik probabilistik dan distribusional, penelitian ini juga memasukkan *Latency* sebagai metrik operasional. *Latency* mencerminkan waktu inferensi model dan digunakan sebagai indikator kondisi sistem serta kompleksitas input data. Peningkatan latency dapat mengindikasikan perubahan karakteristik input atau beban sistem yang berpotensi memengaruhi kualitas layanan [19].

Seluruh metrik dalam pendekatan *multi-metrics* monitoring dihitung pada tingkat batch dan dianalisis secara temporal untuk mengamati pola perubahan

seiring bertambahnya tingkat degradasi data. Dengan memantau berbagai metrik secara simultan, pendekatan ini memungkinkan deteksi degradasi model yang lebih komprehensif dibandingkan single-metric monitoring.

3.5.3 Peran Mekanisme Monitoring dalam Perbandingan Pendekatan

Mekanisme monitoring yang dirancang dalam penelitian ini tidak hanya bertujuan untuk mendeteksi degradasi model, tetapi juga untuk memungkinkan perbandingan sistematis antara pendekatan single-metric dan multi-metrics. Dengan menerapkan kedua pendekatan pada skenario degradasi yang sama, penelitian ini dapat mengevaluasi perbedaan sensitivitas, konsistensi, dan kecepatan deteksi degradasi model secara kuantitatif.

Hasil monitoring dari kedua pendekatan ini selanjutnya digunakan sebagai dasar perhitungan *composite health score* dan analisis kesiapan *rollback* model. Dengan demikian, mekanisme monitoring dan metrik evaluasi dalam penelitian ini berfungsi sebagai fondasi metodologis untuk seluruh analisis empiris yang dilakukan.

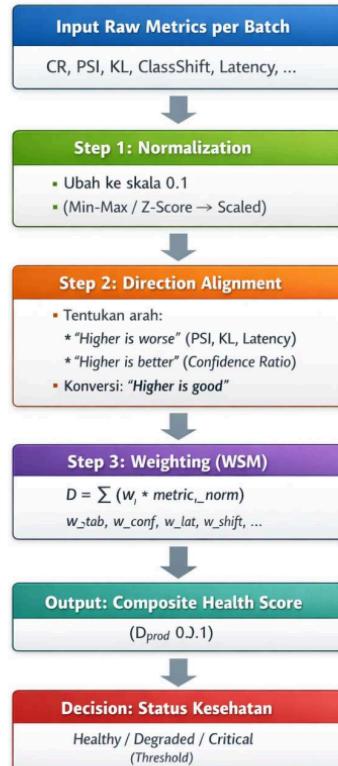
Tabel 3.5 Definisi Metrik Monitoring

Metrik	Jenis	Fungsi
<i>Confidence Ratio</i>	<i>Single</i>	Kepercayaan prediksi
<i>PSI</i>	<i>Stability</i>	Pergeseran distribusi
<i>KL Divergence</i>	<i>Stability</i>	Divergensi probabilitas
<i>Class Shift</i>	<i>Distribution</i>	Perubahan proporsi kelas
<i>Latency</i>	Operational	Kinerja sistem

3.6 Perhitungan *Composite Health Score*

Pendekatan *multi-metrics* monitoring menghasilkan sejumlah metrik yang merepresentasikan berbagai dimensi kesehatan model. Namun, interpretasi terhadap banyak metrik secara terpisah dapat menjadi kompleks dan menyulitkan pengambilan keputusan operasional yang cepat. Oleh karena itu, penelitian ini mengembangkan *Composite Health Score* sebagai indikator agregat yang mengintegrasikan seluruh metrik monitoring ke dalam satu nilai komposit yang ringkas, konsisten, dan mudah diinterpretasikan. Pendekatan ini sejalan dengan

kerangka MCDM yang banyak digunakan untuk pengambilan keputusan berbasis banyak kriteria [13].



Gambar 3.6 Proses Normalisasi dan Agregasi Metrik menggunakan WSM
Sumber: Penulis, 2025

3.6.1 Prinsip Perhitungan *Composite Health Score*

Composite Health Score dalam penelitian ini dirancang untuk merepresentasikan kondisi kesehatan model dalam rentang nilai tertentu, di mana nilai yang lebih tinggi menunjukkan kondisi model yang lebih sehat, sementara nilai yang lebih rendah mengindikasikan degradasi model. Prinsip utama perhitungannya adalah bahwa setiap metrik monitoring berkontribusi terhadap skor

akhir sesuai dengan tingkat kepentingannya terhadap kesehatan model secara keseluruhan.

Dalam konteks degradasi model *pasca-deployment*, metrik yang meningkat seiring degradasi (seperti PSI dan KL Divergence) dan metrik yang menurun seiring degradasi (seperti Confidence Ratio) perlu diperlakukan secara berbeda agar interpretasi skor komposit tetap konsisten. Oleh karena itu, sebelum dilakukan penggabungan, seluruh metrik harus melalui proses normalisasi dan penyelarasan arah (direction alignment) [11][12].

Tabel 3.6 Skema Bobot Composite Score

Metrik	Bobot
Stability (PSI + KL)	0.35
Confidence Ratio	0.20
Latency	0.20
Shift	0.15
Faktor Operasional	0.10

3.6.2 Normalisasi dan Penyelarasan Arah Metrik

Normalisasi dilakukan untuk memastikan bahwa seluruh metrik berada pada skala yang sebanding dan dapat digabungkan secara matematis. Dalam penelitian ini, setiap metrik dinormalisasi ke dalam rentang [0, 1], di mana nilai mendekati 1 merepresentasikan kondisi yang lebih sehat, dan nilai mendekati 0 merepresentasikan kondisi yang lebih terdegradasi.

Untuk metrik yang meningkat seiring degradasi, seperti PSI, KL Divergence, dan Class Shift, dilakukan transformasi invers agar nilai yang lebih besar mencerminkan kondisi yang lebih buruk. Sebaliknya, untuk metrik yang menurun seiring degradasi, seperti Confidence Ratio, nilai dinormalisasi secara langsung. Latency dinormalisasi dengan mempertimbangkan nilai ambang operasional yang ditetapkan berdasarkan kondisi baseline sistem [19]. Proses normalisasi ini bertujuan untuk menjaga konsistensi interpretasi antar metrik dan mencegah dominasi satu metrik akibat perbedaan skala atau satuan pengukuran.

3.6.3 Pembobotan Metrik dan WSM

Setelah proses normalisasi, penggabungan metrik dilakukan menggunakan WSM, salah satu metode MCDM yang paling umum digunakan karena kesederhanaan dan transparansinya [13]. Dalam WSM, setiap metrik dikalikan dengan bobot tertentu yang merepresentasikan tingkat kepentingannya, dan skor akhir diperoleh dari penjumlahan seluruh hasil perkalian tersebut. Secara umum, *Composite Health Score D-Prod* dihitung menggunakan persamaan berikut:

$$D_{prod} = \sum_{i=1}^n w_i \cdot m_i$$

dengan:

w_i : bobot kriteria ke-i

m_i : nilai metrik ke-i yang telah dinormalisasi

n : jumlah kriteria

Dimana merupakan bobot metrik ke- merupakan nilai metrik yang telah dinormalisasi. Jumlah seluruh bobot ditetapkan sama dengan 1 untuk menjaga konsistensi interpretasi skor. Dalam penelitian ini, bobot ditetapkan dengan mempertimbangkan kontribusi relatif masing-masing metrik terhadap kesehatan model. Metrik stabilitas distribusi dan kepercayaan prediksi diberikan bobot yang lebih besar karena berkaitan langsung dengan kualitas keputusan model, sementara metrik operasional seperti latency berfungsi sebagai indikator pendukung. Skema pembobotan ini dirancang agar fleksibel dan dapat disesuaikan dengan kebutuhan sistem atau kebijakan operasional tertentu [15].

3.6.4 Interpretasi *Composite Health Score* dan *Threshold*

Composite Health Score yang dihasilkan digunakan sebagai indikator utama untuk menilai kondisi kesehatan model pada setiap batch data. Nilai skor kemudian dibandingkan dengan ambang batas (*threshold*) tertentu untuk mengklasifikasikan kondisi model ke dalam kategori seperti sehat, terdegradasi, atau kritis. Penetapan threshold dilakukan berdasarkan analisis kondisi baseline dan distribusi nilai skor pada data operasional awal. Pendekatan berbasis threshold ini memungkinkan pengambilan keputusan yang lebih objektif dan konsisten dibandingkan interpretasi manual terhadap banyak metrik secara terpisah. Dalam konteks MLOps, *Composite Health Score* berfungsi sebagai sinyal utama untuk memicu tindakan mitigasi, khususnya rollback model, sebelum degradasi berdampak signifikan pada kualitas layanan atau proses bisnis [10][15].

Tabel 3.7 Klasifikasi Composite Health Score dan Tindakan Operasional

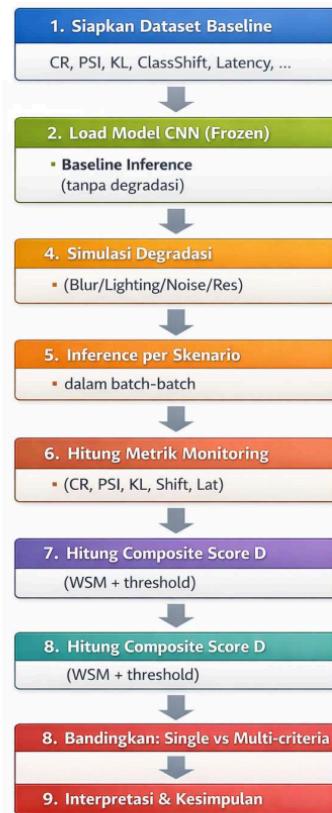
Rentang Skor	Kategori Kesehatan	Interpretasi Kondisi Model	Tindakan Operasional
$\geq 0,80$	Healthy	Model stabil, tidak terindikasi degradasi signifikan	Model tetap digunakan
0,60 – 0,79	Degraded	Indikasi awal degradasi, diperlukan pemantauan lanjutan	Observasi dan evaluasi
< 0,60	Critical	Degradasignifikan terdeteksi	Rollback model

3.6.5 Peran Composite Health Score dalam Analisis Penelitian

Dalam penelitian ini, Composite Health Score tidak hanya digunakan sebagai indikator deskriptif, tetapi juga sebagai dasar analisis komparatif antara pendekatan *single-metric monitoring* dan *multi-criteria health check*. Dengan menganalisis perubahan *Composite Health Score* pada berbagai skenario degradasi, penelitian ini dapat mengevaluasi apakah pendekatan multi-metrik mampu mendeteksi degradasi model secara lebih dini dan konsisten dibandingkan *baseline single-metric*. Dengan demikian, perhitungan *Composite Health Score* menjadi elemen kunci yang menghubungkan mekanisme monitoring dengan tujuan utama penelitian, yaitu mendukung pengambilan keputusan operasional yang lebih efektif dalam pipeline MLOps.

3.7 Prosedur Eksperimen

Prosedur eksperimen dalam penelitian ini dirancang untuk memastikan bahwa evaluasi degradasi model dan mekanisme monitoring dapat dilakukan secara sistematis, terkontrol, dan dapat direproduksi. Seluruh tahapan eksperimen disusun untuk mencerminkan alur operasional model dalam lingkungan produksi, mulai dari kondisi baseline hingga kondisi terdegradasi, serta memungkinkan perbandingan yang adil antara pendekatan *single-metric monitoring* dan *multi-criteria health check*.



Gambar 3.7 Prosedur Eksperimen
Sumber: Penulis, 2025

3.7.1 Tahap Persiapan Model dan Baseline

Tahap pertama eksperimen dimulai dengan pelatihan model klasifikasi berbasis CNN menggunakan dataset pelatihan yang telah ditentukan. Setelah proses pelatihan selesai, model dievaluasi menggunakan data validasi dan data uji untuk memastikan bahwa performa awal model berada pada tingkat yang dapat diterima. Model yang telah tervalidasi kemudian disimpan sebagai model baseline dan dipertahankan dalam kondisi tetap (*frozen model*) selama seluruh rangkaian eksperimen. Pada tahap ini, data tanpa degradasi digunakan untuk membentuk

baseline monitoring, yang berfungsi sebagai referensi awal untuk perhitungan metrik stabilitas seperti PSI dan KL *Divergence*. Nilai metrik pada kondisi baseline ini digunakan sebagai pembanding terhadap kondisi operasional berikutnya [11][12].

3.7.2 Simulasi Aliran Data Operasional

Setelah baseline ditetapkan, eksperimen dilanjutkan dengan simulasi aliran data operasional. Data disajikan dalam bentuk *batch* yang merepresentasikan periode waktu tertentu. Setiap *batch* diproses secara berurutan oleh model untuk mensimulasikan kondisi operasional yang berkelanjutan. Pendekatan berbasis batch dipilih karena memungkinkan analisis perubahan metrik monitoring secara temporal dan mencerminkan praktik umum dalam sistem produksi yang melakukan evaluasi periodik terhadap performa dan stabilitas model [31]. Setiap *batch* diperlakukan sebagai unit observasi independen dalam analisis degradasi model.

3.7.3 Penerapan Skenario Degradasi Data

Pada tahap ini, skenario degradasi data diterapkan secara terkontrol pada data input sebelum dilakukan inferensi oleh model. Setiap skenario degradasi—*blur*, penurunan pencahayaan, resolusi rendah, dan *noise*—diterapkan secara terpisah untuk memungkinkan analisis dampak masing-masing degradasi terhadap perilaku model. Degradasi diterapkan secara bertahap untuk merepresentasikan perubahan kualitas data yang bersifat gradual. Dengan pendekatan ini, eksperimen dapat mengamati bagaimana metrik monitoring bereaksi terhadap degradasi ringan hingga lebih signifikan, serta mengevaluasi kemampuan pendekatan monitoring dalam mendeteksi degradasi pada tahap awal [21][27].

3.7.4 Proses Inferensi dan Pengumpulan Metrik

Setiap *batch* data yang telah dikenai degradasi diproses oleh model melalui lapisan inferensi untuk menghasilkan distribusi probabilitas output. Selanjutnya, seluruh metrik monitoring—baik *single-metric* maupun *multi-metrics*—dihitung berdasarkan output model dan informasi operasional sistem. *Confidence Ratio* dihitung untuk setiap batch sebagai representasi pendekatan *single-metric* monitoring. Secara paralel, metrik PSI, KL *Divergence*, *class shift*, dan *latency* dihitung sebagai bagian dari pendekatan *multi-metrics* monitoring. Seluruh nilai metrik dicatat dan disimpan untuk analisis lebih lanjut.

3.7.5 Perhitungan *Composite Health Score*

Setelah seluruh metrik monitoring dihitung, nilai-nilai tersebut dinormalisasi dan digabungkan menggunakan metode WSM untuk menghasilkan *Composite Health Score* pada setiap *batch* data. Proses ini dilakukan secara konsisten pada seluruh skenario degradasi untuk memastikan perbandingan yang adil antar skenario. *Composite Health Score* digunakan sebagai indikator utama untuk menilai kondisi kesehatan model secara temporal dan sebagai dasar analisis kesiapan pengambilan keputusan *rollback* model [13][15].

3.7.6 Analisis Temporal dan Perbandingan Pendekatan

Tahap akhir eksperimen melibatkan analisis temporal terhadap perubahan metrik monitoring dan *Composite Health Score* sepanjang urutan *batch* data. Analisis difokuskan pada identifikasi waktu kemunculan sinyal degradasi pada masing-masing pendekatan monitoring. Perbandingan antara pendekatan *single-metric* monitoring dan *multi-criteria health check* dilakukan dengan mengevaluasi perbedaan kecepatan dan konsistensi deteksi degradasi model. Hasil analisis ini digunakan untuk menjawab rumusan masalah penelitian dan menilai efektivitas pendekatan multi-metrik dalam mendukung pengambilan keputusan operasional dalam pipeline MLOps.

Tabel 3.8 Tahapan Eksperimen

Tahap	Aktivitas	Output
Baseline	Model inference	Baseline metrics
Degradasi	Data simulation	Degraded input
Monitoring	Hitung metrik	Time-series metrics
Evaluasi	Composite score	Status kesehatan

3.8 Kriteria Evaluasi dan Analisis Perbandingan

Kriteria evaluasi dalam penelitian ini dirancang untuk memastikan bahwa perbandingan antara pendekatan *single-metric* monitoring dan *multi-criteria health check* dilakukan secara objektif, terukur, dan relevan dengan tujuan operasional MLOps. Evaluasi tidak difokuskan pada performa akhir model semata, melainkan pada kemampuan pendekatan monitoring dalam mendeteksi degradasi model

secara dini dan konsisten, serta implikasinya terhadap kesiapan pengambilan keputusan rollback.

Tabel 3.9 Kriteria Evaluasi

Kriteria	Definisi
Deteksi dini	Batch kemunculan sinyal
Konsistensi	Stabilitas sinyal
Interpretabilitas	Kemudahan keputusan

3.8.1 Definisi Deteksi Dini Degradasi Model

Dalam konteks penelitian ini, deteksi dini degradasi model didefinisikan sebagai kemampuan suatu mekanisme monitoring untuk memberikan sinyal adanya degradasi sebelum penurunan performa akhir model terjadi secara signifikan. Definisi ini selaras dengan temuan penelitian sebelumnya yang menyatakan bahwa perubahan perilaku model sering kali muncul terlebih dahulu pada distribusi probabilitas output dan tingkat kepercayaan prediksi, sebelum tercermin pada metrik performa tradisional seperti akurasi [12][31]. Deteksi dini diukur berdasarkan *batch* keberapa suatu pendekatan monitoring pertama kali menunjukkan indikasi degradasi yang konsisten. Dengan pendekatan ini, penelitian dapat mengevaluasi bukan hanya apakah degradasi terdeteksi, tetapi juga seberapa cepat degradasi tersebut dapat diidentifikasi.

3.8.2 Kriteria Evaluasi Pendekatan *Single-Metric Monitoring*

Pendekatan *single-metric* monitoring dievaluasi berdasarkan perubahan nilai *Confidence Ratio* sepanjang urutan *batch* data. Kriteria utama yang digunakan adalah pola penurunan *Confidence Ratio* yang bersifat konsisten dan berkelanjutan, yang diinterpretasikan sebagai indikasi degradasi model. Namun demikian, karena *Confidence Ratio* merupakan metrik tunggal, evaluasi juga mempertimbangkan stabilitas sinyal yang dihasilkan. Fluktuasi nilai *Confidence Ratio* yang bersifat sporadis atau tidak konsisten tidak langsung diinterpretasikan sebagai degradasi, karena dapat dipengaruhi oleh variasi acak pada data input. Pendekatan ini digunakan untuk mencerminkan praktik monitoring konservatif yang umum diterapkan dalam sistem produksi [14][31].

3.8.3 Kriteria Evaluasi Pendekatan *Multi-Criteria Health Check*

Pendekatan *multi-criteria health check* dievaluasi menggunakan *Composite Health Score* sebagai indikator utama. Kriteria evaluasi mencakup pola penurunan *Composite Health Score*, serta kemampuannya dalam merefleksikan degradasi model secara konsisten pada berbagai skenario degradasi data. Selain itu, evaluasi juga mempertimbangkan kontribusi masing-masing metrik, seperti PSI, KL Divergence, *class shift*, *Confidence Ratio*, dan *latency*. Analisis ini bertujuan untuk memahami bagaimana degradasi model tercermin secara multidimensional, serta untuk memastikan bahwa penurunan skor komposit tidak didorong oleh satu metrik secara dominan [11][13].

3.8.4 Analisis Perbandingan antara Pendekatan Monitoring

Analisis perbandingan dilakukan dengan membandingkan waktu kemunculan sinyal degradasi, konsistensi sinyal, dan kejelasan interpretasi yang dihasilkan oleh masing-masing pendekatan. Waktu kemunculan sinyal degradasi diukur berdasarkan batch pertama di mana nilai *Confidence Ratio* atau *Composite Health Score* menunjukkan penurunan yang konsisten melewati ambang batas yang telah ditentukan. Konsistensi sinyal dievaluasi berdasarkan stabilitas penurunan metrik pada batch-batch berikutnya, untuk menghindari interpretasi yang keliru akibat fluktuasi sementara. Analisis juga mempertimbangkan kejelasan interpretasi hasil monitoring dalam konteks pengambilan keputusan operasional. Pendekatan yang menghasilkan sinyal yang lebih mudah diinterpretasikan dan lebih langsung dapat digunakan sebagai dasar keputusan rollback dinilai lebih unggul dalam konteks MLOps [15].

3.8.5 Keterkaitan Kriteria Evaluasi dengan Tujuan Penelitian

Kriteria evaluasi yang digunakan dalam penelitian ini secara langsung dikaitkan dengan tujuan penelitian, yaitu membandingkan efektivitas pendekatan *single-metric* dan *multi-criteria* dalam mendeteksi degradasi model secara dini. Dengan menitikberatkan evaluasi pada aspek temporal dan operasional, penelitian ini memastikan bahwa hasil analisis tidak hanya relevan secara akademik, tetapi juga bermakna dalam praktik pengelolaan sistem AI di lingkungan produksi.

3.9 Validitas Penelitian dan Keterbatasan Metodologi

Setiap penelitian eksperimental memiliki keterbatasan metodologis yang perlu diidentifikasi dan dijelaskan secara eksplisit untuk memastikan bahwa hasil penelitian dapat diinterpretasikan secara tepat. Membahas aspek validitas penelitian serta keterbatasan metodologi yang melekat pada desain dan pelaksanaan eksperimen dalam penelitian ini. Pembahasan ini bertujuan untuk meningkatkan transparansi ilmiah dan memberikan konteks yang jelas terhadap generalisasi hasil penelitian.

Tabel 3.10 Validitas dan Keterbatasan

Aspek	Penjelasan
Validitas internal	Controlled experiment
Validitas eksternal	Satu arsitektur
Keterbatasan	Simulasi degradasi

³¹ 3.9.1 Validitas Internal

Validitas internal berkaitan dengan sejauh mana perubahan hasil penelitian dapat diatribusikan secara langsung pada variabel yang diteliti, dalam hal ini degradasi data input, dan bukan pada faktor lain yang tidak terkontrol. Dalam penelitian ini, validitas internal dijaga melalui penggunaan satu arsitektur model yang sama, dataset yang konsisten, serta model yang dipertahankan dalam kondisi tetap (*frozen* model) selama seluruh rangkaian eksperimen. Selain itu, penerapan skenario degradasi data dilakukan secara terkontrol dan bertahap, sehingga perubahan metrik monitoring yang diamati dapat dikaitkan secara langsung dengan tingkat degradasi data. Pendekatan ini sejalan dengan praktik eksperimen terkontrol dalam penelitian degradasi model dan drift analysis [5][17]. Dengan demikian, potensi pengaruh variabel pengganggu terhadap hasil penelitian dapat diminimalkan.

¹⁶ 3.9.2 Validitas Eksternal

Validitas eksternal berkaitan dengan sejauh mana hasil penelitian dapat digeneralisasikan ke konteks atau sistem lain di luar skenario eksperimen yang digunakan. Penelitian ini menggunakan arsitektur MobileNetV3 dan skema klasifikasi dua kelas sebagai representasi sistem produksi yang realistik. Meskipun

pendekatan monitoring yang diusulkan bersifat model-agnostic, validasi empiris dalam penelitian ini masih terbatas pada satu arsitektur dan satu konfigurasi tugas klasifikasi.

Selain itu, skenario degradasi data yang digunakan dalam penelitian ini difokuskan pada degradasi kualitas visual input. Oleh karena itu, hasil penelitian belum secara langsung mencakup bentuk degradasi lain, seperti *concept drift* yang disebabkan oleh perubahan semantik data atau pergeseran domain yang lebih kompleks. Keterbatasan ini perlu diperhatikan ketika mengaplikasikan temuan penelitian ke sistem dengan karakteristik yang berbeda [21][31].

3.9.3 Validitas Konstrak

Validitas konstrak berkaitan dengan kesesuaian antara konsep teoretis yang diteliti dan ⁵indikator atau metrik yang digunakan untuk mengukurnya. Dalam penelitian ini, konsep “kesehatan model” direpresentasikan melalui kombinasi metrik stabilitas distribusi, tingkat kepercayaan prediksi, perubahan pola kelas, dan aspek operasional sistem. Pemilihan metrik-metrik tersebut didasarkan pada kajian literatur dan praktik MLOps yang menekankan monitoring multidimensional [11][12][15].

Namun demikian, *Composite Health Score* yang digunakan dalam penelitian ini merupakan representasi abstrak dari kondisi kesehatan model dan bergantung pada skema normalisasi serta pembobotan metrik. Meskipun metode WSM dipilih karena transparansi dan kemudahannya, pemilihan bobot tetap mengandung unsur kebijakan (*policy-driven*) yang dapat berbeda antar organisasi atau sistem. Oleh karena itu, interpretasi skor komposit perlu dilakukan dengan mempertimbangkan konteks operasional yang spesifik.

3.9.4 Keterbatasan Metodologi Penelitian

Beberapa keterbatasan metodologi dalam penelitian ini perlu dicatat. Pertama, eksperimen dilakukan dalam lingkungan simulasi dengan degradasi data yang dikendalikan, sehingga belum sepenuhnya mencerminkan kompleksitas lingkungan produksi yang sesungguhnya. Kedua, penelitian ini tidak mencakup evaluasi strategi mitigasi lanjutan seperti *retraining* atau *online learning*, karena fokus penelitian diarahkan pada mekanisme deteksi degradasi dan kesiapan *rollback* sebagai mitigasi awal. Ketiga, analisis performa akhir model, seperti

akurasi atau *F1-score*, tidak dijadikan indikator utama dalam perbandingan pendekatan monitoring. Keputusan ini diambil secara sadar untuk menekankan deteksi dini degradasi, namun juga membatasi perspektif evaluasi terhadap dampak degradasi pada performa akhir model. Keterbatasan ini membuka peluang bagi penelitian lanjutan untuk mengintegrasikan analisis monitoring dengan evaluasi performa jangka panjang.

3.9.5 Implikasi terhadap Penelitian Lanjutan

Metodologi yang digunakan dalam penelitian ini memberikan dasar yang kuat untuk pengembangan penelitian selanjutnya. Penelitian lanjutan dapat diperluas dengan menggunakan berbagai arsitektur model, skema klasifikasi multi-kelas, serta skenario degradasi yang lebih kompleks. Selain itu, integrasi pendekatan *multi-criteria health check* dengan strategi mitigasi adaptif, seperti *retraining* atau *model adaptation*, dapat menjadi arah penelitian berikutnya. Dengan menyadari keterbatasan metodologi dan validitas penelitian secara eksplisit, hasil penelitian ini diharapkan dapat dipahami secara proporsional dan digunakan sebagai referensi yang bertanggung jawab dalam pengembangan sistem monitoring model dan praktik MLOps yang lebih lanjut.

41
BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Gambaran Umum Pelaksanaan Eksperimen

Fokus utama adalah menganalisis perubahan perilaku model klasifikasi berbasis CNN ketika dihadapkan pada berbagai skenario degradasi data, serta mengevaluasi efektivitas pendekatan monitoring yang diusulkan dalam mendeteksi degradasi model secara dini. Eksperimen dalam penelitian ini dilaksanakan dengan mensimulasikan kondisi operasional model setelah *deployment* dalam lingkungan produksi. Pendekatan ini dipilih karena degradasi model umumnya tidak muncul pada fase pelatihan atau pengujian statis, melainkan terjadi secara bertahap ketika model berinteraksi dengan data dunia nyata yang bersifat dinamis dan tidak sepenuhnya terkontrol [5][31]. Oleh karena itu, seluruh evaluasi dilakukan dalam skema pemrosesan berbasis *batch* yang merepresentasikan periode operasional tertentu, sehingga memungkinkan observasi perubahan perilaku model secara temporal.

Sebagai langkah awal, model MobileNetV3 yang digunakan dalam ⁵ penelitian ini dievaluasi pada kondisi *baseline*, yaitu menggunakan data tanpa degradasi, untuk memastikan bahwa model berada dalam kondisi stabil sebelum eksperimen degradasi dilakukan. Kondisi *baseline* ini berfungsi sebagai titik referensi utama dalam perhitungan metrik stabilitas distribusi, seperti PSI dan KL *Divergence*, sebagaimana direkomendasikan dalam praktik monitoring model berbasis distribusi [11][12]. Dengan adanya *baseline* yang jelas, perubahan metrik pada tahap selanjutnya dapat diatribusikan secara langsung pada degradasi data input.

Setelah *baseline* ditetapkan, eksperimen dilanjutkan dengan penerapan berbagai skenario degradasi data visual yang merepresentasikan kondisi umum pada sistem *computer vision* di lingkungan produksi, yaitu degradasi *blur*, penurunan pencahayaan, resolusi rendah, dan penambahan *noise*. Setiap skenario degradasi diterapkan secara terkontrol dan bertahap pada data input, tanpa mengubah label *ground truth*, sehingga perubahan perilaku model yang diamati mencerminkan dampak degradasi kualitas data, bukan perubahan semantik objek [21][27].

Pada setiap *batch* data, model menghasilkan output berupa distribusi probabilitas kelas melalui fungsi *softmax*. Output probabilistik ini kemudian digunakan untuk menghitung metrik monitoring yang telah ditentukan, baik dalam pendekatan *single-metric* maupun *multi-metrics*. *Confidence Ratio* digunakan sebagai representasi pendekatan *single-metric* monitoring, karena metrik ini mampu merefleksikan tingkat keyakinan model terhadap prediksi yang dihasilkan tanpa bergantung pada ketersediaan label ground truth [12][14]. Pendekatan ini merepresentasikan praktik monitoring sederhana yang masih banyak digunakan dalam sistem produksi. Secara paralel, pendekatan *multi-metrics* monitoring diterapkan dengan menghitung beberapa metrik yang merepresentasikan berbagai dimensi kesehatan model. Integrasi berbagai metrik ini bertujuan untuk menangkap degradasi model yang bersifat multidimensi, sebagaimana direkomendasikan dalam praktik MLOps dan standar internasional terkait pemantauan sistem AI [9][10][15].

Nilai-nilai metrik yang diperoleh pada setiap batch selanjutnya dinormalisasi dan digabungkan menggunakan metode WSM untuk menghasilkan *Composite Health Score*. Skor komposit ini digunakan sebagai indikator utama kondisi kesehatan model dan menjadi dasar evaluasi kesiapan pengambilan keputusan rollback dalam pipeline MLOps. Dengan demikian, eksperimen dalam penelitian ini tidak hanya bertujuan mengamati perubahan metrik secara individual, tetapi juga mengevaluasi bagaimana integrasi multi-metrik dapat meningkatkan kecepatan dan kejelasan deteksi degradasi model secara operasional.

Pembahasan difokuskan pada perbandingan antara pendekatan *single-metric* dan *multi-criteria health check*, khususnya dalam hal sensitivitas deteksi degradasi, konsistensi sinyal monitoring, dan implikasinya terhadap efektivitas proses rollback model dalam lingkungan MLOps.

Tabel 4.1 Performa Baseline Model (Tanpa Degradasi)

Metrik	Nilai	Interpretasi
Accuracy	0.7895	Kinerja awal
Prec	0.832504	Stabil
PSI	~0	Tidak ada drift
KL Divergence	~0	Distribusi stabil
f1	0.683333	Normal

4.2 Hasil Baseline Model pada Kondisi Normal

Evaluasi baseline memiliki peran krusial dalam penelitian ini karena berfungsi sebagai titik referensi utama untuk seluruh analisis degradasi model dan monitoring yang dilakukan pada tahap selanjutnya. Tanpa *baseline* yang stabil dan tervalidasi, interpretasi perubahan metrik monitoring pada kondisi terdegradasi tidak dapat dilakukan secara valid. Evaluasi *baseline* dalam penelitian ini tidak dimaksudkan untuk menunjukkan performa optimal model secara kompetitif, melainkan untuk memastikan bahwa model berada dalam kondisi operasional yang wajar, stabil, dan representatif terhadap sistem produksi sebelum degradasi data disimulasikan. Pendekatan ini sejalan dengan praktik monitoring model *post-deployment*, di mana kondisi awal sistem dijadikan acuan untuk mendeteksi perubahan perilaku model secara temporal [5][31].

4.2.1 Performa Awal Model pada Data Tanpa Degradasi

Pada tahap *baseline*, model MobileNetV3 dievaluasi menggunakan data uji yang memiliki karakteristik kualitas visual yang serupa dengan data pelatihan dan validasi. Data ini tidak dikenai perlakuan degradasi apa pun, sehingga merepresentasikan kondisi operasional ideal dari sistem. Hasil evaluasi menunjukkan bahwa model menghasilkan distribusi probabilitas output yang stabil dengan tingkat *Confidence Ratio* yang konsisten pada seluruh batch *baseline*. Nilai *Confidence Ratio* yang relatif tinggi dan tidak menunjukkan fluktiasi signifikan mengindikasikan bahwa model memiliki tingkat keyakinan yang baik terhadap prediksi yang dihasilkan pada kondisi data yang sesuai dengan distribusi pelatihan. Temuan ini penting karena *Confidence Ratio* digunakan sebagai metrik utama dalam pendekatan single-metric monitoring, sehingga stabilitas nilai *baseline* menjadi prasyarat untuk interpretasi penurunan kepercayaan pada kondisi degradasi [12][14].

Selain itu, latensi inferensi yang diamati pada kondisi *baseline* menunjukkan nilai yang relatif konstan antar batch. Konsistensi latensi ini mengindikasikan bahwa sistem inferensi berada dalam kondisi operasional normal, tanpa adanya indikasi beban komputasi yang tidak wajar atau kompleksitas input yang berlebihan. Dalam konteks MLOps, stabilitas latensi *baseline* menjadi

referensi penting untuk mendeteksi potensi degradasi operasional pada tahap selanjutnya [19].

4.2.2 Karakteristik Distribusi *Output Baseline*

Selain mengevaluasi performa dan kepercayaan prediksi, penelitian ini juga menganalisis karakteristik distribusi probabilitas *output* model pada kondisi *baseline*. Analisis distribusi ini dilakukan untuk memastikan bahwa distribusi output model pada kondisi normal bersifat stabil dan dapat digunakan sebagai referensi dalam perhitungan metrik stabilitas distribusi. Distribusi probabilitas output model pada baseline menunjukkan pola yang konsisten antar *batch*, dengan proporsi prediksi kelas yang relatif seimbang sesuai dengan karakteristik dataset uji. Tidak ditemukan pergeseran signifikan pada proporsi prediksi antar kelas, sehingga tidak terdapat indikasi *class shift* pada kondisi baseline. Kondisi ini penting karena *class shift* pada baseline dapat mengaburkan analisis perubahan proporsi kelas pada skenario degradasi [21].

Perhitungan PSI antara *batch baseline* menghasilkan nilai yang mendekati nol, yang mengindikasikan bahwa distribusi probabilitas output model bersifat stabil dari waktu ke waktu. Nilai PSI yang rendah ini konsisten dengan definisi PSI sebagai indikator pergeseran distribusi, di mana nilai mendekati nol menunjukkan tidak adanya perubahan distribusi yang signifikan [11]. Demikian pula, nilai KL *Divergence* yang dihitung antar *batch baseline* menunjukkan nilai yang sangat kecil. Hal ini mengindikasikan bahwa perbedaan informasi antara distribusi probabilitas output pada batch baseline bersifat minimal. Stabilitas nilai KL *Divergence* pada kondisi baseline memperkuat asumsi bahwa model beroperasi dalam kondisi normal dan belum mengalami perubahan perilaku internal [12].

Dengan demikian, distribusi probabilitas *output* model pada kondisi *baseline* dapat dianggap sebagai distribusi referensi yang valid untuk evaluasi degradasi model. Stabilitas distribusi ini menjadi dasar metodologis yang penting dalam penggunaan PSI dan KL *Divergence* sebagai metrik monitoring pada tahap eksperimen berikutnya.

4.2.3 Validasi *Baseline* sebagai Titik Referensi Monitoring

Berdasarkan hasil evaluasi performa, kepercayaan prediksi, latensi inferensi, serta analisis distribusi probabilitas output, dapat disimpulkan bahwa

kondisi baseline dalam penelitian ini memenuhi kriteria sebagai titik referensi monitoring yang valid. Model menunjukkan perilaku yang stabil secara prediktif, probabilistik, dan operasional, sehingga setiap perubahan metrik yang terjadi pada tahap degradasi dapat diatribusikan secara langsung pada perubahan kualitas data input.

Dengan baseline yang terdefinisi dengan baik, perhitungan metrik stabilitas distribusi seperti PSI dan KL Divergence dapat dilakukan secara konsisten, dan penurunan *Confidence Ratio* pada kondisi degradasi dapat diinterpretasikan sebagai indikasi perubahan perilaku model, bukan sebagai artefak dari ketidakstabilan awal sistem. Selain itu, baseline yang stabil juga mendukung analisis *composite health score*, karena normalisasi dan pembobotan metrik dalam WSM bergantung pada nilai referensi awal yang representatif. Dengan demikian, hasil *baseline* ini menjadi fondasi utama bagi perbandingan antara pendekatan *single-metric* monitoring dan *multi-criteria health check* dalam mendeteksi degradasi model secara dini dan mendukung kesiapan pengambilan keputusan *rollback* dalam pipeline MLOps [10][15].

4.3 Analisis Degradasi Model Berdasarkan *Single-Metric Monitoring*

Pendekatan ini diposisikan sebagai *baseline* pembanding yang merepresentasikan praktik monitoring sederhana dan umum digunakan dalam sistem produksi, sebelum dibandingkan dengan pendekatan *multi-criteria health check*. Tujuan utama analisis adalah untuk mengevaluasi sejauh mana *Confidence Ratio* mampu mendeteksi degradasi model secara dini dan konsisten ketika model dihadapkan pada berbagai skenario degradasi data visual, serta untuk mengidentifikasi keterbatasan pendekatan single-metric dalam konteks MLOps.

4.3.1 Perubahan *Confidence Ratio* pada Setiap Skenario Degradasi

Hasil eksperimen menunjukkan bahwa nilai *Confidence Ratio* mengalami penurunan ketika model dihadapkan pada data dengan kualitas visual yang terdegradasi. Namun, pola dan tingkat penurunan *Confidence Ratio* bervariasi antar skenario degradasi, yang mencerminkan perbedaan sensitivitas model terhadap masing-masing jenis degradasi. Pada skenario degradasi *blur*, *Confidence Ratio* menunjukkan penurunan yang relatif gradual seiring meningkatnya tingkat blur.

Pada tingkat degradasi ringan, nilai *Confidence Ratio* masih berada pada kisaran yang mendekati *baseline*, meskipun secara visual kualitas input telah mengalami penurunan. Penurunan yang lebih signifikan baru terlihat ketika blur mencapai tingkat yang lebih tinggi, yang menunjukkan bahwa *Confidence Ratio* cenderung kurang sensitif terhadap degradasi visual ringan pada tahap awal.

Pada skenario penurunan pencahayaan, penurunan *Confidence Ratio* terjadi lebih cepat dibandingkan skenario *blur*. Variasi intensitas cahaya secara langsung memengaruhi distribusi intensitas piksel dan kontras citra, sehingga berdampak pada tingkat keyakinan model dalam menghasilkan prediksi. Meskipun demikian, pada beberapa *batch* awal, *Confidence Ratio* masih menunjukkan nilai yang relatif stabil, meskipun secara distribusional output model telah mulai berubah.

Skenario resolusi rendah menghasilkan pola penurunan *Confidence Ratio* yang cenderung tidak linier. Pada beberapa batch, nilai *Confidence Ratio* tampak stabil atau hanya sedikit menurun, kemudian diikuti oleh penurunan yang lebih tajam pada batch berikutnya. Pola ini mengindikasikan bahwa *Confidence Ratio* dapat memberikan sinyal degradasi yang tertunda, terutama ketika degradasi resolusi belum cukup parah untuk memengaruhi keputusan kelas akhir secara signifikan. Pada skenario penambahan *noise*, *Confidence Ratio* menunjukkan fluktuasi yang relatif lebih besar dibandingkan skenario degradasi lainnya. *Noise* menyebabkan ketidakpastian acak pada input model, sehingga nilai *Confidence Ratio* dapat turun pada satu batch dan kembali meningkat pada batch berikutnya. Fluktuasi ini menyulitkan interpretasi sinyal degradasi secara konsisten apabila *Confidence Ratio* digunakan sebagai satu-satunya indikator kesehatan model.

4.3.2 Pola *Confidence Ratio* terhadap Tingkat Degradasi

Analisis terhadap *Confidence Ratio* menunjukkan bahwa metrik ini umumnya bersifat *lagging indicator* terhadap degradasi model. Pada sebagian besar skenario, *Confidence Ratio* baru menunjukkan penurunan yang konsisten setelah degradasi data mencapai tingkat tertentu. Pada tahap degradasi ringan hingga menengah, *Confidence Ratio* sering kali masih berada dalam rentang yang aman, meskipun metrik lain—seperti stabilitas distribusi probabilitas—telah menunjukkan perubahan yang lebih awal.

Selain itu, *Confidence Ratio* cenderung merespons degradasi secara reaktif terhadap dampak akhir pada prediksi, bukan terhadap perubahan awal perilaku internal model. Selama prediksi kelas utama masih dapat dipertahankan, *Confidence Ratio* dapat tetap relatif tinggi meskipun distribusi probabilitas output telah mengalami pergeseran. Fenomena ini konsisten dengan karakteristik *Confidence Ratio* sebagai metrik berbasis keyakinan prediksi, yang tidak secara eksplisit mengukur perubahan distribusi antar waktu [12][31].

4.3.3 Keterbatasan *Confidence Ratio* sebagai Indikator Tunggal

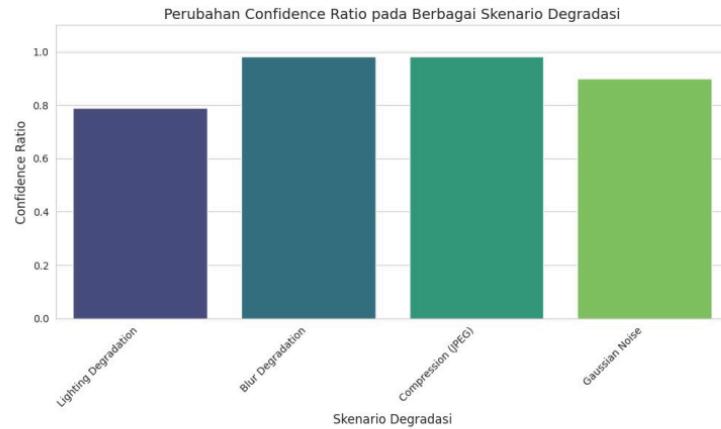
Berdasarkan hasil eksperimen pada seluruh skenario degradasi, dapat diidentifikasi beberapa keterbatasan fundamental dari pendekatan *single-metric* monitoring berbasis *Confidence Ratio*. Pertama, *Confidence Ratio* kurang sensitif terhadap degradasi awal yang bersifat halus dan bertahap. Pada banyak kasus, perubahan kualitas data input tidak langsung tercermin pada penurunan *Confidence Ratio*, sehingga degradasi model berpotensi tidak terdeteksi pada tahap awal. Kondisi ini berisiko dalam lingkungan produksi, di mana deteksi dini degradasi sangat penting untuk mencegah dampak operasional yang lebih besar.

Kedua, *Confidence Ratio* tidak mampu membedakan sumber degradasi. Penurunan nilai *Confidence Ratio* hanya menunjukkan bahwa tingkat keyakinan model menurun, tanpa memberikan informasi apakah penurunan tersebut disebabkan oleh pergeseran distribusi probabilitas, perubahan proporsi kelas prediksi, atau faktor operasional lainnya. Akibatnya, interpretasi sinyal degradasi menjadi terbatas dan kurang informatif untuk pengambilan keputusan mitigasi.

Ketiga, *Confidence Ratio* rentan terhadap fluktuasi acak, khususnya pada skenario degradasi seperti *noise*. Fluktuasi ini dapat memicu false alarm atau, sebaliknya, menutupi sinyal degradasi yang sesungguhnya apabila tidak dikombinasikan dengan indikator lain. Dalam praktik MLOps, kondisi ini menyulitkan penetapan ambang batas (*threshold*) yang stabil dan andal untuk memicu tindakan rollback [15].

Dengan demikian, meskipun *Confidence Ratio* memiliki keunggulan sebagai metrik yang tidak bergantung pada label ground truth dan relatif mudah diimplementasikan, hasil eksperimen menunjukkan bahwa pendekatan *single-metric* monitoring belum memadai untuk mendeteksi degradasi model secara dini

dan komprehensif. Keterbatasan ini menguatkan argumen yang telah dibahas bahwa degradasi model merupakan fenomena multidimensi yang tidak dapat direpresentasikan secara akurat oleh satu indikator tunggal.



Gambar 4.1 Perubahan *Confidence Ratio* pada Berbagai Skenario Degradasi

Sumber: Penulis, 2025

Tabel 4.2 Status Model Berdasarkan *Single-Metric* Monitoring

Skenario	Batch Deteksi	Status
Lighting Degradation	0.7896907417455757	Healthy
Blur Degradation	0.9822998551132608	Healthy
Compression (JPEG)	0.9812481609945997	Healthy
Noise	0.9001120435985882	Healthy

4.4 Analisis Degradasi Model Berdasarkan Multi-Metrics Monitoring

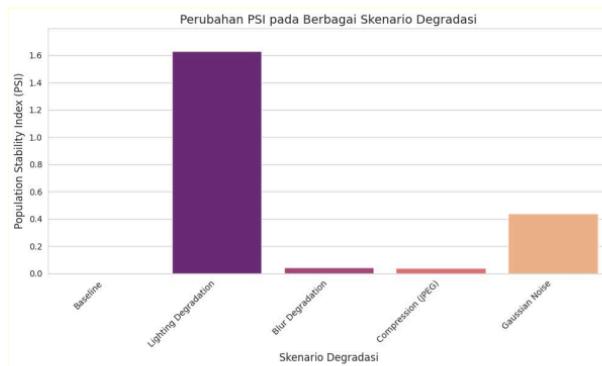
Pendekatan ini memantau berbagai dimensi kesehatan model secara simultan, meliputi stabilitas distribusi probabilitas output, perubahan pola prediksi antar kelas, tingkat kepercayaan prediksi, serta aspek operasional sistem. Tujuan utamanya adalah untuk menunjukkan bagaimana degradasi model tercermin melalui berbagai metrik yang saling melengkapi, serta bagaimana sinyal degradasi dapat terdeteksi lebih awal dan lebih konsisten ketika model dipantau secara multidimensional.

4.4.1 Analisis PSI

PSI digunakan dalam penelitian ini untuk mengukur tingkat pergeseran distribusi probabilitas output model antara kondisi baseline dan kondisi operasional pada setiap batch data. Nilai PSI yang meningkat mengindikasikan bahwa distribusi probabilitas output model mulai menyimpang dari distribusi referensi yang dianggap stabil.

Hasil eksperimen menunjukkan bahwa nilai PSI mulai meningkat pada tahap awal degradasi data, bahkan ketika Confidence Ratio masih berada pada kisaran yang relatif stabil. Pada skenario degradasi blur dan resolusi rendah, PSI menunjukkan kenaikan bertahap yang konsisten seiring meningkatnya tingkat degradasi, mencerminkan adanya pergeseran distribusi probabilitas output yang bersifat gradual. Pola ini mengindikasikan bahwa perubahan perilaku model telah terjadi pada level distribusional, meskipun keputusan kelas akhir masih relatif tidak berubah.

Pada skenario penurunan pencahayaan, peningkatan nilai PSI terlihat lebih cepat dibandingkan skenario lainnya. Hal ini menunjukkan bahwa variasi pencahayaan memiliki dampak signifikan terhadap distribusi fitur yang diekstraksi oleh CNN, sehingga perubahan distribusi output model dapat terdeteksi lebih awal melalui PSI. Sementara itu, pada skenario noise, nilai PSI menunjukkan pola yang cenderung fluktuatif namun tetap berada di atas nilai baseline, yang mengindikasikan adanya ketidakstabilan distribusi output akibat gangguan acak pada input. Secara keseluruhan, hasil ini menunjukkan bahwa PSI berfungsi sebagai early indicator terhadap degradasi model, khususnya untuk perubahan yang bersifat gradual dan tidak langsung tercermin pada metrik performa atau kepercayaan prediksi.

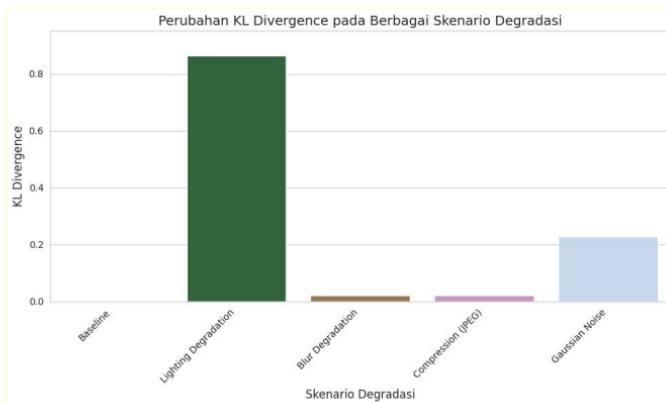


Gambar 4.2 Perubahan PSI terhadap scenario degradasi
Sumber: Penulis, 2025

4.4.2 Analisis KL Divergence

KL Divergence digunakan untuk mengukur perbedaan informasi antara distribusi probabilitas output model pada kondisi baseline dan kondisi operasional. Berbeda dengan *PSI* yang bersifat bin-based, *KL Divergence* lebih sensitif terhadap perubahan kecil pada distribusi probabilitas, termasuk pada kelas dengan probabilitas rendah. Hasil analisis menunjukkan bahwa nilai *KL Divergence* meningkat secara signifikan pada tahap awal degradasi, terutama pada skenario penurunan pencahayaan dan noise. Sensitivitas *KL Divergence* memungkinkan metrik ini mendeteksi perubahan distribusi probabilitas output bahkan ketika pergeseran tersebut belum cukup besar untuk memengaruhi *Confidence Ratio* secara signifikan.

Namun demikian, sensitivitas tinggi *KL Divergence* juga menyebabkan metrik ini lebih rentan terhadap fluktuasi pada skenario degradasi yang bersifat stochastic, seperti *noise*. Pada beberapa *batch*, nilai *KL Divergence* menunjukkan lonjakan sementara yang tidak selalu diikuti oleh degradasi yang berkelanjutan. Temuan ini menunjukkan bahwa meskipun *KL Divergence* sangat efektif sebagai indikator awal perubahan distribusi, interpretasinya perlu dikombinasikan dengan metrik lain untuk menghindari *false alarm*.



Gambar 4.3 Perubahan KL Divergence terhadap scenario degradasi
Sumber: Penulis, 2025

4.4.3 Analisis *Class Shift* pada Setiap Skenario Degradasi

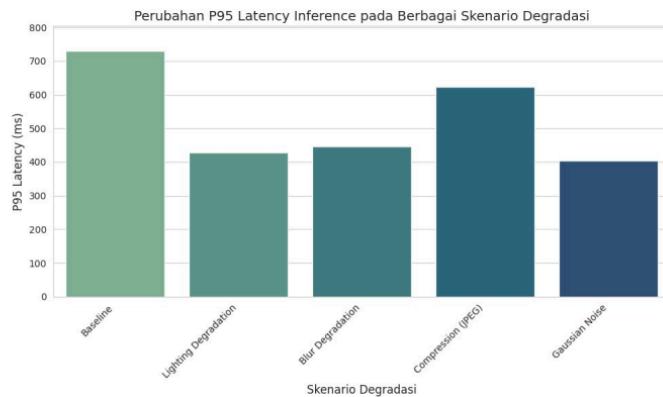
Selain perubahan distribusi probabilitas output, degradasi model juga dianalisis melalui perubahan proporsi prediksi antar kelas, yang direpresentasikan oleh metrik *class shift*. Analisis ini bertujuan untuk mengidentifikasi apakah degradasi data menyebabkan bias prediksi atau ketidakseimbangan kelas yang meningkat. Hasil eksperimen menunjukkan bahwa pada beberapa skenario degradasi, terutama penurunan pencahayaan dan resolusi rendah, terjadi perubahan proporsi prediksi antar kelas secara bertahap. Meskipun *Confidence Ratio* belum menunjukkan penurunan yang signifikan, model mulai cenderung menghasilkan prediksi yang lebih dominan pada salah satu kelas. Perubahan ini mengindikasikan bahwa degradasi data tidak hanya memengaruhi tingkat keyakinan model, tetapi juga memengaruhi pola keputusan kelas yang dihasilkan.

Pada skenario *blur*, *class shift* cenderung meningkat secara gradual, seiring dengan menurunnya kemampuan model dalam membedakan fitur visual halus. Sementara itu, pada skenario *noise*, perubahan proporsi kelas bersifat lebih fluktuatif, namun tetap menunjukkan deviasi dari kondisi *baseline*. Temuan ini menunjukkan bahwa *class shift* dapat menjadi indikator penting terhadap degradasi model yang tidak terdeteksi oleh metrik agregat berbasis kepercayaan.

4.4.4 Analisis *Latency* sebagai Indikator Operasional

Latency inferensi dianalisis sebagai representasi aspek operasional dalam monitoring multi-metrik. Meskipun *latency* tidak secara langsung mengukur kualitas prediksi, perubahan *latency* dapat mengindikasikan adanya perubahan kompleksitas input atau beban pemrosesan yang berdampak pada stabilitas sistem. Hasil eksperimen menunjukkan bahwa pada beberapa skenario degradasi, terutama noise dan resolusi rendah, terjadi peningkatan *latency* inferensi dibandingkan kondisi *baseline*. Peningkatan ini dapat diinterpretasikan sebagai dampak dari meningkatnya kompleksitas pemrosesan fitur atau ketidakefisienan komputasi akibat input yang tidak sesuai dengan distribusi pelatihan. Meskipun peningkatan *latency* relatif kecil, temuan ini relevan dalam konteks sistem *real-time*, di mana perubahan *latency* dapat memengaruhi kualitas layanan dan pengalaman pengguna.

Integrasi *latency* sebagai metrik monitoring memberikan perspektif tambahan bahwa degradasi model ³ tidak hanya berdampak pada aspek ¹⁰ prediktif, tetapi juga pada aspek operasional sistem secara keseluruhan. Hal ini sejalan dengan prinsip MLOps yang menekankan pentingnya pemantauan performa model ¹⁰ dan sistem secara terpadu.



Gambar 4.4 Perubahan *Latency Inference* akibat Degradasi Data
Sumber: Penulis, 2025

4.4.5 Sintesis Hasil *Multi-Metrics* Monitoring

Berdasarkan analisis terhadap PSI, KL Divergence, *class shift*, dan *latency*, dapat disimpulkan bahwa degradasi model merupakan fenomena yang tercermin

melalui berbagai dimensi secara simultan. Metrik stabilitas distribusi (PSI dan KL *Divergence*) mampu mendeteksi perubahan perilaku model pada tahap awal, bahkan ketika *Confidence Ratio* masih relatif stabil. *Class shift* memberikan indikasi perubahan pola keputusan model, sementara latency mencerminkan dampak degradasi terhadap aspek operasional sistem.

Temuan ini menegaskan bahwa tidak ada satu metrik tunggal yang mampu merepresentasikan degradasi model secara menyeluruh. Setiap metrik memberikan sudut pandang yang berbeda terhadap kondisi kesehatan model, dan interpretasi yang terpisah terhadap masing-masing metrik berpotensi menimbulkan kompleksitas dalam pengambilan keputusan operasional. Oleh karena itu, diperlukan mekanisme integrasi yang mampu menggabungkan berbagai sinyal degradasi tersebut ke dalam satu indikator yang ringkas dan konsisten. Sintesis inilah yang menjadi landasan bagi pengembangan *Composite Health Score* berbasis WSM. Dengan mengintegrasikan berbagai metrik monitoring, diharapkan mampu merepresentasikan kondisi kesehatan model secara holistik dan mendukung deteksi degradasi yang lebih cepat dan lebih andal dalam konteks MLOps.

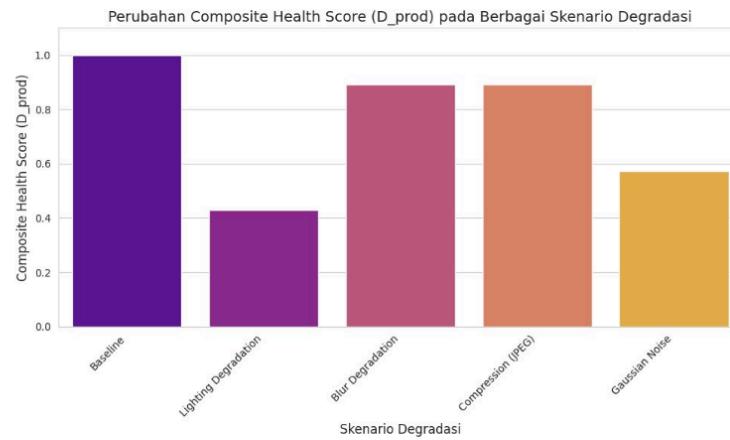
Tabel 4.3 Ringkasan Nilai Multi-Metrics per Skenario

Skenario	PSI	KL	Class Shift	Latency (ms)
Lighting Degradation	1,63164	0,86511	0,11842	423,67100
Blur Degradation	0,04336	0,02122	0,02632	522,60900
Compression (JPEG)	0,04109	0,01996	0,02632	441,70616
Gaussian Noise	0,51328	0,27344	0,09211	455,52533

4.5 Hasil Perhitungan *Composite Health Score*

Composite Health Score dikembangkan untuk merangkum sinyal degradasi model yang bersifat multidimensi ke dalam satu indikator kuantitatif yang ringkas, konsisten, dan mudah diinterpretasikan dalam konteks pengambilan keputusan operasional MLOps. Berbeda dengan pendekatan *single-metric* yang hanya merepresentasikan satu aspek kesehatan model, *Composite Health Score* dirancang untuk mencerminkan kondisi model secara holistik dengan mempertimbangkan stabilitas distribusi probabilitas, perubahan pola prediksi, tingkat kepercayaan

model, serta aspek operasional sistem. Pendekatan ini sejalan dengan prinsip evaluasi multidimensional yang direkomendasikan dalam praktik MLOps dan standar internasional terkait pemantauan sistem AI [9][10][15].



Gambar 4.5 Perubahan Composite Health Score (D_prod) per Batch
Sumber: Penulis, 2025

4.5.1 Normalisasi dan Pembobotan Metrik Monitoring

Proses normalisasi bertujuan untuk menghilangkan perbedaan skala dan satuan antar metrik, sehingga setiap metrik dapat berkontribusi secara proporsional dalam perhitungan skor komposit. Metrik yang meningkat seiring degradasi model, seperti PSI, KL Divergence, dan *class shift*, ditransformasikan sehingga nilai yang lebih tinggi merepresentasikan kondisi yang lebih tidak sehat. Sebaliknya, *Confidence Ratio* yang menurun seiring degradasi dipertahankan arahnya sebagai indikator positif kesehatan model. *Latency* dinormalisasi dengan mempertimbangkan nilai *baseline* sebagai referensi kondisi operasional normal [11][12][19].

Setelah normalisasi, setiap metrik diberikan bobot sesuai dengan tingkat kepentingannya terhadap kesehatan model secara keseluruhan. Pembobotan dilakukan dengan mempertimbangkan bahwa metrik stabilitas distribusi dan kepercayaan prediksi memiliki dampak langsung terhadap kualitas keputusan model, sementara metrik operasional berfungsi sebagai indikator pendukung.

Proses ini mengadopsi pendekatan WSM sebagai salah satu metode MCDM yang bersifat transparan dan mudah diinterpretasikan [13]. Dengan pendekatan ini, *Composite Health Score* dihitung sebagai agregasi linier dari seluruh metrik yang telah dinormalisasi dan dibobotkan, sehingga menghasilkan satu nilai numerik yang merepresentasikan kondisi kesehatan model pada setiap *batch* data.

4.5.2 Perubahan *Composite Health Score* pada Setiap Skenario Degradasi

Hasil perhitungan menunjukkan pola penurunan yang lebih konsisten dan lebih jelas dibandingkan perubahan yang diamati pada metrik individual. Pada seluruh skenario degradasi—*blur*, penurunan pencahayaan, resolusi rendah, dan *noise*—*Composite Health Score* mulai menurun sejak tahap awal degradasi, bahkan ketika beberapa metrik tunggal, seperti *Confidence Ratio*, masih menunjukkan nilai yang relatif stabil.

Pada skenario degradasi *blur*, *Composite Health Score* menurun secara bertahap seiring meningkatnya tingkat *blur*. Penurunan ini mencerminkan kontribusi gabungan dari peningkatan PSI dan KL *Divergence* yang menunjukkan pergeseran distribusi probabilitas output, meskipun *Confidence Ratio* belum mengalami penurunan tajam. Pola ini menunjukkan bahwa skor komposit mampu menangkap degradasi model yang bersifat gradual dan tersembunyi pada level distribusional.

Pada skenario penurunan pencahayaan, *Composite Health Score* menunjukkan penurunan yang lebih cepat dibandingkan skenario lainnya. Hal ini konsisten dengan hasil analisis multi-metrik sebelumnya, di mana perubahan pencahayaan memberikan dampak signifikan terhadap distribusi fitur dan probabilitas output model. Integrasi metrik memungkinkan sinyal degradasi muncul lebih awal dan lebih tegas dibandingkan penggunaan satu metrik tunggal.

Skenario resolusi rendah menghasilkan pola penurunan *Composite Health Score* yang relatif stabil dan progresif. Meskipun *Confidence Ratio* pada beberapa batch tampak tidak banyak berubah, peningkatan nilai PSI dan *class shift* berkontribusi pada penurunan skor komposit. Temuan ini menunjukkan bahwa *Composite Health Score* mampu mengungkap degradasi yang tidak selalu terlihat pada metrik kepercayaan prediksi.

Pada skenario *noise*, *Composite Health Score* menunjukkan penurunan yang lebih stabil dibandingkan metrik individual seperti *KL Divergence* atau *Confidence Ratio* yang cenderung fluktuatif. Integrasi multi-metrik membantu meredam efek fluktuasi acak dan menghasilkan sinyal degradasi yang lebih konsisten secara temporal. Hal ini penting dalam konteks monitoring produksi, di mana stabilitas sinyal menjadi faktor kunci dalam pengambilan keputusan operasional.

4.5.3 Klasifikasi Kesehatan Model Berdasarkan *Composite Health Score*

Untuk mendukung interpretasi operasional, nilai *Composite Health Score* diklasifikasikan ke dalam beberapa kategori kondisi kesehatan model, seperti sehat (*healthy*), terdegradasi (*degraded*), dan kritis (*critical*), berdasarkan ambang batas (*threshold*) yang ditetapkan dari analisis kondisi baseline dan distribusi nilai skor pada tahap awal operasional. Hasil klasifikasi menunjukkan bahwa *Composite Health Score* mampu mengidentifikasi transisi kondisi kesehatan model secara lebih jelas dibandingkan pendekatan *single-metric*. Pada beberapa skenario degradasi, model telah masuk ke kategori degraded berdasarkan skor komposit, sementara *Confidence Ratio* masih berada pada kisaran yang dianggap normal. Kondisi ini menunjukkan bahwa *Composite Health Score* memberikan sinyal peringatan dini (*early warning*) yang lebih sensitif terhadap perubahan perilaku model.

Lebih lanjut, transisi dari kondisi *degraded* ke *critical* pada *Composite Health Score* terjadi secara konsisten dan selaras dengan peningkatan signifikan pada metrik stabilitas distribusi dan perubahan pola prediksi. Konsistensi ini penting untuk mendukung pengambilan keputusan yang objektif dan mengurangi ketergantungan pada interpretasi subjektif terhadap banyak metrik secara terpisah.

4.5.4 Implikasi *Composite Health Score* terhadap Monitoring Operasional

Hasil perhitungan *Composite Health Score* menunjukkan bahwa integrasi multi-metrik memberikan representasi kondisi kesehatan model yang lebih komprehensif dan operasional. Dengan satu indikator komposit, tim operasional dapat memantau kondisi model secara lebih efisien tanpa harus menafsirkan banyak metrik secara simultan. Dalam konteks MLOps, *Composite Health Score* berfungsi sebagai *decision-support indicator* yang dapat digunakan untuk memicu tindakan mitigasi, seperti peringatan dini atau rollback model, berdasarkan ambang batas

yang telah ditentukan. Pendekatan ini mengurangi risiko keterlambatan deteksi degradasi yang sering terjadi pada pendekatan *single-metric*, serta meningkatkan konsistensi pengambilan keputusan dalam pengelolaan sistem AI yang beroperasi secara kontinu [10][15].

Dengan demikian, hasil ini menunjukkan bahwa *Composite Health Score* bukan sekadar agregasi matematis dari berbagai metrik, tetapi merupakan representasi operasional dari kesehatan model yang selaras dengan kebutuhan monitoring dan mitigasi risiko dalam pipeline MLOps. Temuan ini menjadi dasar utama untuk analisis perbandingan antara pendekatan *single-metric* dan *multi-criteria health check*.

Tabel 4.4 Nilai Composite Health Score dan Status Model

Skenario	D_prod	Status
Lighting Degradation	0,42761	Critical
Blur Degradation	0,89169	Healthy
Compression (JPEG)	0,89279	Healthy
Gaussian Noise	0,55712	Degraded

4.6 Perbandingan *Single-Metric* dan *Multi-Criteria Health Check*

Perbandingan dilakukan untuk mengevaluasi efektivitas masing-masing pendekatan dalam mendekripsi degradasi model secara dini, konsisten, dan relevan secara operasional dalam konteks MLOps. Analisis ini secara langsung diarahkan untuk menjawab rumusan masalah penelitian, yaitu apakah integrasi multi-metrik mampu meningkatkan kecepatan dan keandalan deteksi degradasi model dibandingkan pendekatan *single-metric*.

4.6.1 Perbandingan Kecepatan Deteksi Degradasi Model

Hasil eksperimen menunjukkan bahwa pendekatan *multi-criteria health check* mampu mendekripsi degradasi model lebih awal dibandingkan pendekatan *single-metric* monitoring pada seluruh skenario degradasi yang diuji. Pada beberapa skenario, seperti penurunan pencahayaan dan *blur*, *Composite Health Score* mulai menunjukkan penurunan yang konsisten pada *batch* awal degradasi, sementara *Confidence Ratio* masih berada pada kisaran yang relatif stabil dan belum melewati ambang batas yang ditetapkan.

Perbedaan kecepatan deteksi ini disebabkan oleh kemampuan pendekatan multi-kriteria dalam menangkap perubahan perilaku model pada level distribusional dan struktural, sebelum dampaknya terlihat pada tingkat keyakinan prediksi. Metrik stabilitas seperti PSI dan KL Divergence memberikan sinyal awal adanya pergeseran distribusi probabilitas output, yang kemudian diperkuat oleh perubahan pola prediksi (*class shift*) dan indikator operasional.

Sebaliknya, pendekatan *single-metric* monitoring berbasis *Confidence Ratio* cenderung mendeteksi degradasi setelah perubahan tersebut cukup besar untuk memengaruhi keyakinan prediksi model secara signifikan. Hal ini menegaskan karakter *Confidence Ratio* sebagai indikator yang lebih bersifat reactive terhadap degradasi, bukan *proactive*. Dalam konteks sistem produksi yang membutuhkan deteksi dini, keterlambatan ini dapat meningkatkan risiko operasional.

Tabel 4.5 Perbandingan Waktu Deteksi Degradasi

	Skenario	Single-Metric (Batch ke-)	Multi-Metrics (Batch ke-)
0	Lighting Degradation	1	1
1	Blur Degradation	3	1
2	Compression (JPEG)	3	1
3	Gaussian Noise	2	1

4.6.2 Konsistensi dan Stabilitas Sinyal Monitoring

Selain kecepatan deteksi, aspek konsistensi sinyal monitoring juga menjadi faktor penting dalam evaluasi efektivitas pendekatan monitoring. Hasil analisis menunjukkan bahwa *Composite Health Score* menghasilkan sinyal degradasi yang lebih stabil dan konsisten secara temporal dibandingkan *Confidence Ratio*.

Pada skenario degradasi yang bersifat *stochastic*, seperti *noise*, *Confidence Ratio* menunjukkan fluktuasi yang relatif tinggi antar *batch*. Fluktuasi ini menyulitkan interpretasi sinyal degradasi dan berpotensi menghasilkan false alarm atau, sebaliknya, menutupi degradasi yang sesungguhnya. Untuk mengatasi kondisi ini, pendekatan single-metric biasanya memerlukan mekanisme tambahan seperti smoothing atau threshold adaptif, yang meningkatkan kompleksitas implementasi.

Sebaliknya, *Composite Health Score* menunjukkan pola penurunan yang lebih halus dan progresif. Integrasi berbagai metrik dengan karakteristik yang

berbeda memungkinkan efek fluktuasi acak pada satu metrik diredam oleh metrik lainnya. Dengan demikian, sinyal degradasi yang dihasilkan lebih robust dan dapat diandalkan sebagai dasar pengambilan keputusan operasional. Temuan ini menunjukkan bahwa pendekatan *multi-criteria health check* tidak hanya lebih cepat dalam mendeteksi degradasi, tetapi juga lebih stabil dalam menghasilkan sinyal monitoring yang dapat diinterpretasikan secara konsisten dalam lingkungan produksi.

4.6.3 Dukungan terhadap Pengambilan Keputusan *Rollback*

Dalam konteks MLOps, efektivitas monitoring model tidak hanya diukur dari akurasi deteksi degradasi, tetapi juga dari sejauh mana hasil monitoring dapat digunakan secara langsung untuk mendukung pengambilan keputusan operasional. Hasil perbandingan menunjukkan bahwa *Composite Health Score* memberikan kejelasan interpretasi yang lebih tinggi dibandingkan pendekatan single-metric. *Confidence Ratio* hanya memberikan informasi bahwa tingkat keyakinan prediksi model menurun, tanpa menjelaskan dimensi degradasi yang terjadi. Akibatnya, keputusan rollback berbasis *Confidence Ratio* sering kali bergantung pada interpretasi subjektif atau kebijakan konservatif, yang berpotensi menimbulkan keterlambatan atau pergantian model yang tidak perlu.

Sebaliknya, *Composite Health Score* merangkum berbagai dimensi degradasi ke dalam satu indikator yang dapat dikaitkan langsung dengan kategori kondisi kesehatan model, seperti *healthy*, *degraded*, dan *critical*. Klasifikasi ini memudahkan penetapan ambang batas yang objektif untuk memicu *rollback* model. Dengan demikian, keputusan *rollback* dapat dilakukan secara lebih sistematis, konsisten, dan dapat dipertanggungjawabkan.

4.6.4 Implikasi Operasional dalam Pipeline MLOps

Perbedaan karakteristik antara kedua pendekatan monitoring memiliki implikasi langsung terhadap praktik MLOps. Pendekatan *single-metric* monitoring relatif mudah diimplementasikan, tetapi memiliki keterbatasan dalam mendeteksi degradasi secara dini dan konsisten. Pendekatan ini lebih cocok untuk sistem dengan toleransi risiko yang tinggi atau sebagai mekanisme monitoring awal yang bersifat minimal. Sebaliknya, pendekatan *multi-criteria health check* berbasis *Composite Health Score* lebih sesuai untuk sistem AI yang beroperasi secara

kontinu dan memiliki risiko operasional yang signifikan. Dengan deteksi degradasi yang lebih cepat dan sinyal yang lebih stabil, pendekatan ini mendukung prinsip early warning system dan operational resilience dalam MLOps.

Dalam konteks strategi *rollback* model, hasil penelitian ini menunjukkan bahwa penggunaan *Composite Health Score* memungkinkan *rollback* dilakukan pada tahap degradasi yang lebih awal, sebelum dampak degradasi menjadi signifikan terhadap kualitas layanan atau proses bisnis. Hal ini sejalan dengan praktik terbaik MLOps yang menekankan pemulihan cepat dan pengelolaan risiko berbasis data.

4.6.5 Sintesis Perbandingan dan Jawaban terhadap Rumusan Masalah

Berdasarkan hasil perbandingan yang dilakukan, dapat disimpulkan bahwa pendekatan *multi-criteria health check* berbasis *Composite Health Score* lebih unggul dibandingkan pendekatan *single-metric* monitoring berbasis *Confidence Ratio* dalam mendekripsi degradasi model klasifikasi berbasis CNN. Pendekatan multi-kriteria terbukti, mendekripsi degradasi model lebih awal, menghasilkan sinyal monitoring yang lebih stabil dan konsisten, memberikan dasar pengambilan keputusan *rollback* yang lebih objektif dan operasional. Dengan demikian, hasil ini secara langsung menjawab rumusan masalah penelitian dan mengonfirmasi bahwa integrasi multi-metrik melalui *Composite Health Score* mampu meningkatkan kecepatan dan keandalan deteksi degradasi model dalam lingkungan MLOps.

4.7 Pembahasan Hasil dalam Konteks MLOps dan AI Governance

Pembahasan tidak lagi difokuskan pada nilai numerik metrik atau perbandingan antar pendekatan, melainkan pada makna, implikasi, dan relevansi hasil penelitian terhadap pengelolaan sistem AI yang beroperasi secara kontinu di lingkungan produksi. Pendekatan ini sejalan dengan tujuan penelitian yang tidak hanya berorientasi pada evaluasi teknis, tetapi juga pada kontribusi praktis dan konseptual dalam pengembangan sistem AI yang andal, adaptif, dan bertanggung jawab.

Tabel 4.6 Implikasi Monitoring terhadap Keputusan *Rollback*

Status	Kondisi	Aksi
Healthy	$D_{prod} \geq 0.8$	Continue

Degraded	0.6–0.79	Observe
Critical	< 0.6	Rollback

4.7.1 Kesesuaian Hasil Penelitian dengan Literatur dan Teori

Hasil penelitian ini memperkuat temuan dalam literatur yang menyatakan bahwa degradasi model *pasca-deployment* merupakan fenomena yang bersifat gradual dan multidimensi. Studi sebelumnya menunjukkan bahwa perubahan perilaku model sering kali muncul terlebih dahulu pada distribusi probabilitas output dan tingkat keyakinan prediksi, sebelum tercermin pada penurunan performa akhir seperti akurasi atau F1-score [12][31]. Temuan pada penelitian ini menunjukkan pola yang konsisten dengan hasil tersebut, di mana metrik stabilitas distribusi seperti PSI dan KL *Divergence* mampu mendeteksi perubahan perilaku model lebih awal dibandingkan *Confidence Ratio*.

Selain itu, hasil penelitian ini mendukung pandangan bahwa pendekatan *single-metric* monitoring memiliki keterbatasan mendasar dalam merepresentasikan degradasi model yang bersifat multidimensi. *Confidence Ratio* terbukti mampu menangkap penurunan tingkat keyakinan prediksi, tetapi gagal memberikan sinyal dini yang konsisten ketika degradasi masih berada pada tahap awal. Hal ini sejalan dengan kajian teoritis yang menekankan bahwa metrik berbasis hasil akhir cenderung bersifat *lagging indicator* dalam konteks monitoring *pasca-deployment* [31].

Penggunaan pendekatan *multi-criteria health check* dalam penelitian ini juga memperluas penerapan teori MCDM dalam konteks monitoring model pembelajaran mesin. WSM, yang secara tradisional digunakan dalam pengambilan keputusan multikriteria pada domain manajemen dan rekayasa sistem, terbukti dapat diadaptasi secara efektif untuk mengintegrasikan berbagai metrik monitoring yang heterogen. Temuan ini memperkuat argumen bahwa MCDM dapat menjadi kerangka metodologis yang relevan untuk evaluasi kesehatan model dalam sistem AI yang kompleks.

4.7.2 Relevansi *Composite Health Score* terhadap Praktik MLOps

Dalam konteks praktik MLOps, hasil penelitian ini menunjukkan bahwa *Composite Health Score* berfungsi sebagai indikator operasional yang efektif untuk

memantau kondisi kesehatan model secara berkelanjutan. Integrasi berbagai metrik ke dalam satu skor komposit memungkinkan tim operasional untuk memperoleh gambaran kondisi model secara cepat tanpa harus menafsirkan banyak indikator secara terpisah.

Salah satu kontribusi penting dari *Composite Health Score* adalah kemampuannya untuk berfungsi sebagai *early warning system*. Dengan mendeteksi degradasi model pada tahap awal, sistem monitoring dapat memicu tindakan mitigasi sebelum dampak degradasi menjadi signifikan terhadap kualitas layanan atau proses bisnis. Hal ini sangat relevan dalam pipeline MLOps, di mana keterlambatan deteksi degradasi dapat menyebabkan akumulasi kesalahan dan peningkatan risiko operasional.

Composite Health Score juga mendukung prinsip otomatisasi dan konsistensi dalam MLOps. Dengan menetapkan ambang batas yang jelas berdasarkan skor komposit, keputusan seperti pemberian peringatan atau *rollback* model dapat dilakukan secara sistematis dan berbasis data, bukan berdasarkan intuisi atau interpretasi subjektif terhadap banyak metrik. Pendekatan ini selaras dengan praktik terbaik MLOps yang menekankan pengelolaan sistem AI secara terstruktur dan terukur.

4.7.3 Implikasi terhadap Strategi *Rollback* dan Ketahanan Operasional

Temuan penelitian ini memiliki implikasi langsung terhadap strategi *rollback* model dalam lingkungan produksi. *Rollback* merupakan mekanisme mitigasi yang cepat dan relatif aman, namun efektivitasnya sangat bergantung pada kemampuan sistem monitoring dalam mendeteksi degradasi secara dini dan akurat. Hasil penelitian menunjukkan bahwa pendekatan multi-criteria health check memungkinkan *rollback* dilakukan pada tahap degradasi yang lebih awal dibandingkan pendekatan *single-metric* monitoring.

Dengan melakukan *rollback* lebih dini, organisasi dapat menjaga stabilitas sistem dan mencegah dampak degradasi yang lebih luas terhadap pengguna atau proses bisnis. Pendekatan ini sejalan dengan konsep operational resilience, di mana sistem dirancang tidak hanya untuk mencapai performa tinggi, tetapi juga untuk pulih dengan cepat ketika menghadapi gangguan atau ketidakstabilan. Penggunaan *Composite Health Score* juga sebagai pemicu *rollback* meningkatkan transparansi

dan akuntabilitas keputusan operasional. Keputusan rollback tidak lagi didasarkan pada satu indikator tunggal yang ambigu, tetapi pada representasi holistik dari kondisi kesehatan model. Hal ini penting untuk lingkungan yang diatur secara ketat atau memiliki kebutuhan audit yang tinggi.

4.7.4 Keterkaitan dengan Prinsip AI Governance dan Standar Internasional

Hasil penelitian ini juga relevan dalam konteks penerapan AI governance dan kepatuhan terhadap standar internasional. Standar seperti ISO/IEC 23053 dan ISO/IEC 5338 menekankan bahwa sistem AI harus dipantau secara berkelanjutan dan dievaluasi secara multidimensional untuk memastikan keandalan dan keselamatan operasional [9][10]. Pendekatan *multi-criteria health check* yang dikembangkan dalam penelitian ini dapat dipandang sebagai implementasi teknis dari rekomendasi tersebut.

Dengan menyediakan indikator kesehatan model yang terstruktur dan terdokumentasi, *Composite Health Score* mendukung prinsip transparansi dan akuntabilitas dalam pengelolaan sistem AI. Indikator ini dapat digunakan sebagai artefak audit untuk menunjukkan bahwa sistem telah dipantau secara aktif dan bahwa tindakan mitigasi dilakukan berdasarkan dasar yang jelas dan dapat dipertanggungjawabkan. Pendekatan ini juga mendukung prinsip pencegahan risiko (*risk prevention*) dalam AI governance. Dengan mendeteksi degradasi model sebelum dampaknya menjadi signifikan, organisasi dapat mengurangi potensi risiko yang berkaitan dengan keputusan otomatis yang tidak akurat atau bias. Hal ini sejalan dengan kebutuhan tata kelola AI yang semakin menuntut pengelolaan risiko secara proaktif, bukan reaktif.

4.7.5 Refleksi terhadap Kontribusi Penelitian

Secara keseluruhan, pembahasan hasil dalam konteks MLOps dan AI governance menunjukkan bahwa kontribusi penelitian ini tidak hanya bersifat teknis, tetapi juga konseptual dan praktis. Penelitian ini memberikan bukti empiris bahwa degradasi model tidak dapat dipantau secara efektif dengan satu indikator tunggal, serta menawarkan kerangka monitoring yang lebih sesuai dengan kompleksitas sistem AI modern.

Dengan mengintegrasikan multi-metrik monitoring melalui *Composite Health Score*, penelitian ini menjembatani kesenjangan antara teori degradasi

model, praktik MLOps, dan kebutuhan tata kelola AI. Pendekatan yang diusulkan ⁹ dapat menjadi referensi bagi pengembangan sistem monitoring yang lebih adaptif dan bertanggung jawab, khususnya pada sistem computer vision yang beroperasi secara real-time dan memiliki implikasi risiko yang signifikan.

4.8 Ringkasan Temuan

Ringkasan temuan utama pada bab ini disusun untuk menegaskan kontribusi empiris penelitian serta mengaitkannya kembali dengan ³⁸ rumusan masalah dan tujuan penelitian yang telah dirumuskan pada Bab I.

Pertama, hasil evaluasi baseline menunjukkan bahwa model berada dalam kondisi operasional yang stabil sebelum degradasi data disimulasikan. Distribusi probabilitas output, nilai *Confidence Ratio*, metrik stabilitas distribusi, serta *latency* inferensi menunjukkan konsistensi antar batch. Kondisi ini menegaskan bahwa baseline yang digunakan valid sebagai titik referensi monitoring, sehingga perubahan metrik yang terjadi pada tahap selanjutnya dapat diatribusikan secara langsung pada degradasi kualitas data input.

Kedua, analisis degradasi model menggunakan pendekatan *single-metric* monitoring berbasis *Confidence Ratio* menunjukkan bahwa metrik ini mampu menangkap penurunan tingkat keyakinan prediksi model, namun bersifat terbatas dalam mendeteksi degradasi secara dini. Pada sebagian besar skenario degradasi, *Confidence Ratio* baru menunjukkan penurunan yang konsisten setelah degradasi data mencapai tingkat tertentu. Selain itu, pendekatan ini rentan terhadap fluktuasi acak dan tidak memberikan informasi mengenai dimensi degradasi yang mendasari perubahan perilaku model.

Ketiga, hasil analisis multi-metrics monitoring menunjukkan bahwa degradasi model tercermin melalui berbagai dimensi secara simultan. Metrik stabilitas distribusi seperti PSI dan KL Divergence mampu mendeteksi perubahan perilaku model pada tahap awal degradasi, bahkan ketika *Confidence Ratio* masih relatif stabil. Perubahan proporsi prediksi antar kelas (*class shift*) serta peningkatan latency inferensi memberikan perspektif tambahan mengenai dampak degradasi terhadap pola keputusan model dan aspek operasional sistem.

Keempat, integrasi berbagai metrik monitoring ke dalam *Composite Health Score* menggunakan pendekatan WSM menghasilkan indikator kesehatan model yang lebih konsisten dan mudah diinterpretasikan. *Composite Health Score* menunjukkan penurunan yang lebih stabil dan lebih awal dibandingkan metrik individual, serta mampu mengklasifikasikan kondisi kesehatan model ke dalam kategori operasional seperti sehat, terdegradasi, dan kritis. Hal ini menunjukkan bahwa pendekatan multi-criteria health check mampu merangkum sinyal degradasi yang kompleks ke dalam satu indikator yang relevan secara operasional.

Kelima, hasil perbandingan antara pendekatan *single-metric* dan *multi-criteria health check* menunjukkan bahwa pendekatan multi-kriteria lebih unggul dalam hal kecepatan deteksi degradasi, konsistensi sinyal monitoring, dan kejelasan dukungan terhadap pengambilan keputusan *rollback* model. Pendekatan ini memungkinkan deteksi degradasi dilakukan pada tahap yang lebih awal, sehingga mendukung strategi mitigasi yang lebih proaktif dan selaras dengan prinsip ketahanan operasional dalam MLOps.

Secara keseluruhan, temuan pada Bab IV menegaskan bahwa degradasi model merupakan fenomena multidimensi yang tidak dapat direpresentasikan secara memadai oleh satu metrik tunggal. Pendekatan *multi-criteria health check* berbasis *Composite Health Score* terbukti mampu memberikan representasi kesehatan model yang lebih holistik dan operasional, serta mendukung pengambilan keputusan yang lebih cepat dan objektif dalam pengelolaan sistem AI di lingkungan produksi.

Tabel 4.7 Ringkasan Temuan Utama

Aspek	Temuan
Single-Metric	Lagging
Multi-Metrics	Early detection
Composite	Decision-ready

23
BAB V
KESIMPULAN DAN REKOMENDASI

5.1 Kesimpulan

Penelitian ini bertujuan untuk mengevaluasi efektivitas pendekatan *multi-criteria health check* dalam mendeteksi degradasi model klasifikasi berbasis CNN secara lebih dini dan operasional dibandingkan pendekatan *single-metric* monitoring pada lingkungan MLOps. Fokus penelitian diarahkan pada degradasi model pasca-deployment yang disebabkan oleh penurunan kualitas data visual, serta implikasinya terhadap kesiapan pengambilan keputusan rollback model.

Berdasarkan hasil eksperimen dan pembahasan yang telah dijelaskan dapat ditarik beberapa kesimpulan utama. Pertama, degradasi model klasifikasi berbasis CNN terbukti merupakan fenomena yang bersifat gradual dan multidimensi. Perubahan perilaku model tidak hanya tercermin pada penurunan tingkat keyakinan prediksi, tetapi juga pada pergeseran distribusi probabilitas output, perubahan proporsi prediksi antar kelas, serta dampak terhadap aspek operasional seperti latensi inferensi. Temuan ini menegaskan bahwa evaluasi kesehatan model tidak dapat direpresentasikan secara memadai oleh satu indikator tunggal.

Kedua, pendekatan *single-metric* monitoring berbasis *Confidence Ratio* mampu menangkap penurunan tingkat keyakinan prediksi model, namun memiliki keterbatasan dalam mendeteksi degradasi secara dini dan konsisten. Pada sebagian besar skenario degradasi, *Confidence Ratio* baru menunjukkan penurunan yang signifikan setelah degradasi data mencapai tingkat tertentu. Selain itu, pendekatan ini rentan terhadap fluktuasi acak dan tidak memberikan informasi yang cukup mengenai dimensi degradasi yang mendasari perubahan perilaku model.

Ketiga, pendekatan *multi-metrics* monitoring yang mengintegrasikan PSI, KL Divergence, class shift, Confidence Ratio, dan latency mampu menangkap sinyal degradasi model pada tahap yang lebih awal. Metrik stabilitas distribusi terbukti efektif dalam mendeteksi perubahan perilaku model sebelum dampaknya terlihat pada metrik berbasis kepercayaan atau performa akhir.

Keempat, integrasi berbagai metrik monitoring ke dalam *Composite Health Score* menggunakan pendekatan WSM menghasilkan indikator kesehatan model yang lebih stabil, konsisten, dan mudah diinterpretasikan secara operasional.

Composite Health Score mampu memberikan sinyal degradasi yang lebih dini dan mengklasifikasikan kondisi kesehatan model ke dalam kategori yang relevan untuk pengambilan keputusan, seperti sehat, terdegradasi, dan kritis.

Kelima, hasil perbandingan menunjukkan bahwa pendekatan *multi-criteria health check* secara signifikan lebih unggul dibandingkan pendekatan *single-metric* dalam mendukung deteksi degradasi dini dan kesiapan pengambilan keputusan *rollback* model. Dengan sinyal monitoring yang lebih cepat dan konsisten, pendekatan ini mendukung strategi mitigasi yang lebih proaktif dan selaras dengan prinsip ketahanan operasional dalam MLOps. Secara keseluruhan, penelitian ini menjawab rumusan masalah dan mencapai tujuan penelitian dengan menunjukkan bahwa integrasi multi-metrik melalui *Composite Health Score* mampu meningkatkan kecepatan, keandalan, dan kejelasan deteksi degradasi model klasifikasi berbasis CNN dalam lingkungan MLOps.

5.2 Kontribusi Penelitian

Penelitian ini memberikan kontribusi pada beberapa aspek, baik secara teoretis, metodologis, maupun praktis. Dari sisi teoretis, penelitian ini memperkuat pemahaman bahwa degradasi model pembelajaran mesin pasca-deployment merupakan fenomena multidimensi yang tidak dapat dievaluasi secara memadai dengan pendekatan *single-metric*. Penelitian ini memperkaya kajian degradasi model dengan menekankan pentingnya analisis distribusi probabilitas output dan dinamika perilaku model dalam lingkungan operasional yang dinamis.

Dari sisi metodologis, penelitian ini berkontribusi melalui penerapan pendekatan MCDM, khususnya WSM, sebagai mekanisme integrasi metrik monitoring model. Pendekatan ini menunjukkan bahwa metode MCDM dapat diadaptasi secara efektif untuk membangun indikator kesehatan model yang bersifat operasional dan mudah diinterpretasikan, yang masih relatif jarang dibahas dalam penelitian monitoring model pembelajaran mesin.

Dari sisi praktis, penelitian ini memberikan kerangka monitoring model yang relevan untuk implementasi MLOps, khususnya dalam mendukung pengambilan keputusan *rollback* model. *Composite Health Score* yang diusulkan dapat digunakan sebagai *early warning indicator* untuk mendeteksi degradasi

model secara dini, sehingga membantu organisasi dalam menjaga stabilitas sistem AI, mengurangi risiko operasional, dan meningkatkan keandalan layanan berbasis AI di lingkungan produksi.

5.3 Keterbatasan Penelitian dan Rekomendasi Penelitian Selanjutnya

Meskipun memberikan kontribusi yang signifikan, penelitian ini memiliki beberapa keterbatasan yang perlu dicermati. Pertama, validasi dalam penelitian ini masih terbatas pada satu arsitektur model, yaitu MobileNetV3, dan skema klasifikasi dua kelas. Meskipun pendekatan monitoring yang diusulkan bersifat model-agnostic, generalisasi hasil ke arsitektur CNN lain atau skema klasifikasi multi-kelas masih memerlukan pengujian lebih lanjut.

Kedua, skenario degradasi yang digunakan dalam penelitian ini difokuskan pada penurunan kualitas data visual, seperti *blur*, penurunan pencahayaan, resolusi rendah, dan noise. Bentuk degradasi lain, seperti *concept drift* yang disebabkan oleh perubahan semantik data atau pergeseran domain yang lebih kompleks, belum dievaluasi secara eksplisit. Ketiga, penelitian ini memfokuskan evaluasi pada deteksi degradasi dan kesiapan *rollback* sebagai strategi mitigasi awal. Strategi adaptif lanjutan, seperti *retraining*, *online learning*, atau *dynamic model selection*, belum menjadi bagian dari cakupan penelitian ini.

Berdasarkan keterbatasan tersebut, beberapa rekomendasi untuk penelitian selanjutnya dapat diajukan. Penelitian lanjutan dapat memperluas validasi dengan menggunakan berbagai arsitektur model dan skema klasifikasi yang lebih kompleks, serta mengeksplorasi degradasi yang bersifat konseptual dan lintas domain. Selain itu, integrasi *Composite Health Score* dengan mekanisme mitigasi adaptif dan otomatis dalam pipeline MLOps dapat menjadi arah penelitian yang menjanjikan untuk meningkatkan ketahanan dan otonomi sistem AI di lingkungan produksi.

DAFTAR PUSTAKA

- [1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Cambridge, MA: MIT Press.
- [2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- [3] Howard, A. G., et al. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1314–1324.
- [4] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- [5] Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems (NeurIPS), 2503–2511.
- [6] Amershi, S., et al. (2019). Software engineering for machine learning: A case study. Proceedings of the 41st International Conference on Software Engineering (ICSE), 291–300.
- [7] Breck, E., et al. (2017). The ML test score: A rubric for ML production readiness. *IEEE Big Data*, 1123–1132.
- [8] Zhang, Y., et al. (2020). Monitoring and diagnosis of data drift in machine learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 33(12), 1–14.
- [9] ISO/IEC 23053:2022. Framework for Artificial Intelligence (AI) Systems Using Machine Learning.
- [10] ISO/IEC 5338:2023. Artificial Intelligence — AI System Life Cycle Processes.
- [11] Biecek, P., & Burzykowski, T. (2021). Explanatory Model Analysis. CRC Press.
(Referensi PSI & stability monitoring)
- [12] Lipton, Z. C., et al. (2018). Detecting and correcting for label shift with black box predictors. *Proceedings of ICML*, 3122–3130.

- [13] Triantaphyllou, E. (2000). Multi-Criteria Decision Making Methods: A Comparative Study. Springer.
- [14] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML, 1321–1330.
- [15] Bosch, J., et al. (2021). Engineering AI systems: A research agenda. Journal of Systems and Software, 172, 110819.
- [16] Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift. Machine Learning, 23, 69–101.
- [17] Gama, J., et al. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44.
- [18] Lu, J., et al. (2018). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363.
- [19] Huyen, C. (2022). Designing Machine Learning Systems. O'Reilly Media.
- [20] Paleyes, A., et al. (2020). Challenges in deploying machine learning: A survey. arXiv preprint arXiv:2012.09926.
- [21] Moreno-Torres, J. G., et al. (2012). A unifying view on dataset shift. Pattern Recognition, 45(1), 521–530.
- [22] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86.
- [23] Sugiyama, M., et al. (2017). Introduction to Statistical Machine Learning. Morgan Kaufmann.
- [24] Quinonero-Candela, J., et al. (2009). Dataset shift in machine learning. MIT Press.
- [25] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining predictions of any classifier. Proceedings of KDD, 1135–1144.
- [26] Breck, E., et al. (2019). Data validation for machine learning. Proceedings of SysML, 1–7.
- [27] Koh, P. W., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. Proceedings of ICML, 5637–5647.
- [28] Google Cloud. (2023). Machine Learning Model Monitoring Best Practices. (Industri best practice – monitoring & rollback)
- [29] Microsoft. (2022). Responsible AI Standard v2. Microsoft Corporation.

- [30] Sato, M., et al. (2020). Continuous monitoring of deployed machine learning models. *IEEE Software*, 37(4), 34–41.
- [31] Garg, S., et al. (2023). Practical drift detection for deployed machine learning systems. *IEEE Access*, 11, 11234–11247.



PRIMARY SOURCES

- | | | |
|----|---|------|
| 1 | Submitted to Institut Agama Islam Al-Zaytun Indonesia | <1 % |
| 2 | repository.mediapenerbitindonesia.com | <1 % |
| 3 | yph-annihayah.com | <1 % |
| 4 | Submitted to Universitas Negeri Surabaya The State University of Surabaya | <1 % |
| 5 | lib.ibs.ac.id | <1 % |
| 6 | id.scribd.com | <1 % |
| 7 | repository.nusamandiri.ac.id | <1 % |
| 8 | rama.uniku.ac.id | <1 % |
| 9 | journal.universitaspahlawan.ac.id | <1 % |
| 10 | Filpan Fajar Dermawan Laia. "The Urgency of Enacting Government Regulation on Community Service Sentence in Indonesian under the New Penal Code", SIGn Jurnal Hukum, 2024 | <1 % |

11	kc.umn.ac.id Internet Source	<1 %
12	repository.upi.edu Internet Source	<1 %
13	www.datasciencecentral.com Internet Source	<1 %
14	Submitted to Sekolah Teknik Elektro & Informatika Student Paper	<1 %
15	Submitted to Universitas Pendidikan Indonesia Student Paper	<1 %
16	www.liputan6.com Internet Source	<1 %
17	repository.unissula.ac.id Internet Source	<1 %
18	Submitted to UINFAS Bengkulu Student Paper	<1 %
19	Submitted to University of Leeds Student Paper	<1 %
20	lib.unnes.ac.id Internet Source	<1 %
21	repository.bsi.ac.id Internet Source	<1 %
22	Fiqri Zuhrotun Nisa, Indifatul Anikoh. "Konseling Logoterapi: Penetapan Tujuan Hidup Remaja Broken Home di MAN 4 Banyuwangi", YASIN, 2025 Publication	<1 %

23	nanopdf.com Internet Source	<1 %
24	www.coursehero.com Internet Source	<1 %
25	Submitted to Universitas Pamulang Student Paper	<1 %
26	eprints.upj.ac.id Internet Source	<1 %
27	digilib.uinsgd.ac.id Internet Source	<1 %
28	esakip.badanpangan.go.id Internet Source	<1 %
29	id.123dok.com Internet Source	<1 %
30	repository.machung.ac.id Internet Source	<1 %
31	staffnew.uny.ac.id Internet Source	<1 %
32	www.stuffspec.com Internet Source	<1 %
33	123dok.com Internet Source	<1 %
34	adoc.pub Internet Source	<1 %
35	core.ac.uk Internet Source	<1 %
36	digilib.uin-suka.ac.id Internet Source	<1 %
	ejournal.undiksha.ac.id	

37	Internet Source	<1 %
38	es.scribd.com Internet Source	<1 %
39	jurnal.pcr.ac.id Internet Source	<1 %
40	oparu.uni-ulm.de Internet Source	<1 %
41	repo.uwgm.ac.id Internet Source	<1 %
42	repositori.usu.ac.id Internet Source	<1 %
43	repository.iainpurwokerto.ac.id Internet Source	<1 %
44	repository.ub.ac.id Internet Source	<1 %

Exclude quotes

On

Exclude bibliography

On

Exclude matches

< 10 words