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



TESIS

TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar

Magister Komputer (M.Kom)

Sony Harianto

14230030

PROGRAM STUDI MAGISTER ILMU KOMPUTER (S2)

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS NUSA MANDIRI

JAKARTA

2025

# SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME

Yang bertanda tangan di bawah ini :

Nama : Sony Harianto

NIM : 14230030

Program Studi : Ilmu Komputer

Fakultas : Teknologi Informasi

Jenjang : Strata Dua (S2)

Peminatan : Software Engineering & Data Science

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: “Peningkatan Akurasi Deteksi Degradasi Model Klasifikasi Berbasis *Multi*-*Criteria* *Decision* *Model*” adalah hasil karya sendiri, dan semua sumber baik yang dikutip maupun yang dirujuk telah saya nyatakan dengan benar, serta belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri dicabut/dibatalkan.

Jakarta, 08 Januari 2026

Sony Harianto

# SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya :

Nama : Sony Harianto

NIM : 14230030

Program Studi : Ilmu Komputer

Fakultas : Teknologi Informasi

Jenjang : Strata Dua (S2)

Peminatan : Software Engineering & Data Science

Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan izin kepada pihak Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri, Hak Bebas Royalti Non-Eksklusif (Non-exclusive Royalti-Free Right) atas karya ilmiah saya yang berjudul : “Peningkatan Akurasi Deteksi Degradasi Model Klasifikasi Berbasis *Multi-Criteria Decision Model”* beserta perangkat yang diperlukan (apabila ada).

Dengan Hak Bebas Royalti Non-Eksklusif ini, pihak Universitas Nusa Mandiri berhak menyimpan, mengalih-media atau bentuk-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 08 Januari 2026

Sony Harianto

# HALAMAN PERSETUJUAN DAN PENGESAHAN

Tesis ini diajukan oleh :

Nama : Sony Harianto

NIM : 14230030

Program Studi : Ilmu Komputer

Fakultas : Teknologi Informasi

Jenjang : Strata Dua (S2)

Peminatan : Software Engineering & Data Science

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

Judul Tesis :

Untuk dipertahankan pada periode 2025-2 dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Komputer (M.Kom) pada Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri.

Jakarta, 08 Januari 2026

PEMBIMBING TESIS

Pembimbing : Dr. Eng. Muhammad Haris ……………...…………..

DEWAN PENGUJI

Penguji I : ……………………………….. …………...……………..

Penguji II : ……………………………….. ………...………………..

Penguji III/ : Dr. Eng. Muhammad Haris ...………………………..

Pembimbing

# LEMBAR BIMBINGAN TESIS

|  |  |
| --- | --- |
| A logo with text on it  AI-generated content may be incorrect. | LEMBAR BIMBINGAN TESIS |
| UNIVERSITAS NUSA MANDRI |

NIM : 14230030

Nama Lengkap : Sony Harianto

Dosen Pembimbing : Dr. Eng. Muhammad Haris

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

Judul Tesis :

|  |  |  |  |
| --- | --- | --- | --- |
| No | Tanggal Bimbingan | Pokok Bahasan | Paraf Dosen Pembimbing |
| 1 | 24/09/2025 | Laporan Proposal dan Dataset |  |
| 2 | 5/10/2025 | Dataset Image |  |
| 3 | 7/10/2025 | Research Question dan Hipotesis |  |
| 4 | 4/11/2025 | Draft Eksperimen |  |
| 5 | 3/12/2025 | Augmentasi Dataset Image |  |
| 6 | 8/12/2025 | Finalisasi Eksperimen |  |
| 7 | 18/12/2025 | Check Laporan Tesis |  |
| 8 | 6/01/2026 | Dokumen Laporan Tesis Final |  |

Catatan untuk dosen pembimbing

Bimbingan Tesis

\* Dimulai pada tanggal : ……………….

\* Diakhiri pada tanggal : ……………….

\* Jumlah pertemuan bimbingan : 8 (delapan) kali bimbingan

Disetujui oleh,

Dosen Pembimbing

(Dr. Eng. Muhammad Haris)

# KATA PENGANTAR

Puji syukur alhamdullillah, penulis panjatkan kehadirat Allah, SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan laporan tesis ini tepat pada waktunya. Dimana laporan tesis ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul tesis, yang penulis ambil sebagai berikut “Peningkatan Akurasi Deteksi Degradasi Model Klasifikasi Berbasis *Multi-Criteria Decision Model*”.

Tujuan penulisan laporan tesis ini dibuat sebagai salah satu syarat untuk Mendapatkan gelar Magister Komputer (M.Kom) pada Program Studi Ilmu Komputer

(S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri.

Penulis juga mencari dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, internet, dll yang terkait dengan pembahasan pada laporan tesis ini.

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak

dalam pembuatan laporan tesis ini, maka penulis tidak dapat menyelesaikan laporan

tesis ini tepat pada waktunya. Untuk itu pada kesempatan ini penulis menyampaikan

ucapan terimakasih kepada :

1. Rektor Universitas Nusa Mandiri
2. Wakil Rektor I Bidang Akademik Universitas Nusa Mandiri
3. Dekan Fakultas Teknologi Informasi Universitas Nusa Mandiri
4. Ketua Program Studi Ilmu Komputer (S2)
5. Bapak Dr. Eng. Muhammad Haris selaku dosen pembimbing yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini
6. Orang tua tercinta yang telah memberikan dukungan material dan moral kepada penulis
7. Istri dan Anak-anak yang selalu memberikan dukungan dan juga memahami situasi dari Suami/Ayahnya.
8. Seluruh Dosen Program Studi Ilmu Komputer (S2) Fakultas Teknologi Informasi Universitas Nusa Mandiri yangtelah memberikan pelajaran yang berarti bagi penulis selama menempuh studi
9. Seluruh staff di lingkungan Universitas Nusa Mandiri yang telah melayani penulis dengan baik selama kuliah.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan laporan tesis ini. Penulis menyadari bahwa penulisan laporan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah di masa yang akan datang.

Akhir kata semoga laporan tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 08 Januari 2026

Sony Harianto

# ABSTRAK

Nama : Sony Harianto

NIM : 14230030

Program Studi : Ilmu Komputer

Fakultas : Teknologi Informasi

Jenjang : Strata Dua (S2)

Peminatan : Software Engineering & Data Science

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

Judul Tesis :

Degradasi model pembelajaran mesin merupakan tantangan penting dalam penerapan sistem klasifikasi berbasis computer vision pada lingkungan produksi. Model yang memiliki performa tinggi pada fase pelatihan berpotensi mengalami penurunan kualitas perilaku ketika dihadapkan pada perubahan karakteristik data operasional, khususnya penurunan kualitas visual input. Pendekatan monitoring model yang masih mengandalkan satu metrik utama, seperti *confidence* *ratio*, memiliki keterbatasan dalam mendeteksi degradasi secara dini karena tidak merepresentasikan kondisi kesehatan model secara menyeluruh.

Penelitian ini mengusulkan pendekatan *multi-criteria decision model* berbasis *composite* *health* *score* untuk mendeteksi degradasi model klasifikasi berbasis *Convolutional Neural Network* (CNN) dalam lingkungan *Machine Learning Operations* (MLOps). Pendekatan ini mengintegrasikan beberapa metrik monitoring, yaitu *Confidence* *Ratio*, *Population* *Stability* *Index* (PSI), *Kullback*–*Leibler* *Divergence* (KL Divergence), *class* *shift*, dan *latency* *inference*, ke dalam satu indikator komposit menggunakan metode *Weighted* *Sum* *Model* (WSM). Eksperimen dilakukan menggunakan arsitektur MobileNetV3 pada skenario klasifikasi dua kelas dengan simulasi degradasi data visual, meliputi blur, penurunan pencahayaan, resolusi rendah, dan noise.

Hasil penelitian menunjukkan bahwa pendekatan *multi-criteria decision model* mampu mendeteksi degradasi model secara lebih cepat dan konsisten dibandingkan pendekatan *single*-*metric* monitoring. *Composite* *health* *score* memberikan indikasi awal degradasi sebelum penurunan metrik kepercayaan prediksi terlihat secara signifikan, sehingga mendukung pengambilan keputusan rollback yang lebih objektif dalam pipeline MLOps. Pendekatan ini diharapkan dapat menjadi dasar pengembangan sistem monitoring model yang lebih adaptif dan andal dalam mendukung ketahanan operasional sistem AI.

Kata kunci: Degradasi model, MLOps, *Multi-criteria Decision Model*, *Composite Health Score*, CNN.

# ABSTRACT

Name : Sony Harianto

NIM : 14230030

Study of Program : Computer Science  
Faculty : Information Technology

Levels : Master Degree (S2)  
Concentration : Software Engineering & Data Science

Improving the Accuracy of Degradation Detection in Classification Models Using a Multi-Criteria Decision Model

Title :

Model degradation is a critical challenge in the deployment of computer vision–based classification systems in production environments. Models that demonstrate high performance during the training phase may experience a gradual decline in behavioral quality when exposed to changes in operational data characteristics, particularly variations in visual input quality. Model monitoring practices that rely on a single primary metric, such as confidence ratio, are limited in their ability to detect early-stage degradation, as they do not adequately represent the overall health of the model.

This research proposes a multi-criteria decision model based on a composite health score to detect degradation in Convolutional Neural Network (CNN)–based classification models within a Machine Learning Operations (MLOps) environment. The proposed approach integrates multiple monitoring metrics—namely Confidence Ratio, Population Stability Index (PSI), Kullback–Leibler (KL) Divergence, class shift, and inference latency—into a single composite indicator using the Weighted Sum Model (WSM) method. Experiments were conducted using the MobileNetV3 architecture for a binary classification task under controlled visual data degradation scenarios, including blur, reduced lighting, low resolution, and noise.

The results demonstrate that the proposed multi-criteria decision model detects model degradation more quickly and consistently than the single-metric monitoring approach. The composite health score provides early indications of degradation before significant declines in prediction confidence are observed, thereby supporting more objective and timely rollback decisions within the MLOps pipeline. This approach is expected to serve as a foundation for developing more adaptive and reliable model monitoring systems that enhance the operational resilience of AI systems.

*Keywords: Model degradation, MLOps, Multi-criteria Decision Model, Composite Health Score, CNN.*

# DAFTAR ISI

[SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME i](#_Toc220220076)

[SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS ii](#_Toc220220077)

[HALAMAN PERSETUJUAN DAN PENGESAHAN iii](#_Toc220220078)

[LEMBAR BIMBINGAN TESIS iv](#_Toc220220079)

[KATA PENGANTAR v](#_Toc220220080)

[ABSTRAK vii](#_Toc220220081)

[ABSTRACT viii](#_Toc220220082)

[DAFTAR ISI ix](#_Toc220220083)

[BAB I 1](#_Toc220220084)

[PENDAHULUAN 1](#_Toc220220085)

[1.1 Latar Belakang 1](#_Toc220220086)

[1.2 Identifikasi Masalah 3](#_Toc220220087)

[1.3 Rumusan Masalah 4](#_Toc220220088)

[1.4 Tujuan Penelitian 5](#_Toc220220089)

[1.5 Ruang Lingkup Penelitian 6](#_Toc220220090)

[1.6 Manfaat Penelitian 8](#_Toc220220091)

[1.7 Hipotesis Penelitian 9](#_Toc220220092)

[1.8 Sistematika Penulisan 10](#_Toc220220093)

[BAB II 12](#_Toc220220094)

[TINJAUAN PUSTAKA 12](#_Toc220220095)

[2.1 *Machine Learning* dan *Computer* *Vision* dalam Lingkungan Produksi 12](#_Toc220220096)

[2.2 Degradasi Model Pembelajaran Mesin di Produksi 14](#_Toc220220097)

[2.3 Concept Drift dan Data Drift 17](#_Toc220220098)

[2.4 Monitoring Model Pembelajaran Mesin 19](#_Toc220220099)

[2.5 Single-Metric Monitoring Berbasis Population Stability Index (PSI) 22](#_Toc220220100)

[2.6 Keterbatasan Pendekatan Single-Metric dalam Monitoring Model 25](#_Toc220220101)

[2.7 Pendekatan Multi-Metric dalam Evaluasi dan Monitoring Model 27](#_Toc220220102)

[2.8 Multi-Criteria Decision Model (MCDM) 29](#_Toc220220103)

[2.9 Weighted Sum Model (WSM) sebagai Metode Agregasi 31](#_Toc220220104)

[2.10 Penelitian Terdahulu 33](#_Toc220220105)

[2.11 Posisi dan Research Gap Penelitian 40](#_Toc220220106)

[BAB III 44](#_Toc220220107)

[METODOLOGI PENELITIAN 44](#_Toc220220108)

[3.1 Desain dan Pendekatan Penelitian 44](#_Toc220220109)

[3.2 Objek dan Lingkup Penelitian 45](#_Toc220220110)

[3.3 Model dan Dataset Penelitian 48](#_Toc220220111)

[3.4 Alur Metode Penelitian 51](#_Toc220220112)

[3.5 Skenario Degradasi Data 53](#_Toc220220113)

[3.6 Mekanisme Monitoring 57](#_Toc220220114)

[3.7 Definisi, Normalisasi, dan Agregasi Metrik 61](#_Toc220220115)

[3.8 Penentuan Threshold Composite Health Score 64](#_Toc220220116)

[3.9 Kriteria Evaluasi dan Analisis Perbandingan 68](#_Toc220220117)

[3.10 Validitas dan Keterbatasan Metodologi 69](#_Toc220220118)

[BAB IV 72](#_Toc220220119)

[HASIL PENELITIAN DAN PEMBAHASAN 72](#_Toc220220120)

[4.1 Gambaran Umum Pelaksanaan Eksperimen 72](#_Toc220220121)

[4.2 Hasil Baseline Model pada Kondisi Normal 74](#_Toc220220122)

[4.3 Hasil Monitoring pada Berbagai Skenario Degradasi 75](#_Toc220220123)

[4.4 Analisis Monitoring Berdasarkan Pendekatan Single-Metric 82](#_Toc220220124)

[4.5 Analisis Monitoring Berdasarkan Pendekatan Multi-Metric 84](#_Toc220220125)

[4.6 Hasil Perhitungan Composite Health Score 88](#_Toc220220126)

[4.7 Penentuan dan Validasi Status Kesehatan Model 89](#_Toc220220127)

[4.8 Perbandingan Pendekatan Single-Metric dan Multi-Metric 90](#_Toc220220128)

[4.9 Pembahasan dalam Konteks MLOps dan Monitoring Produksi 91](#_Toc220220129)

[BAB V 93](#_Toc220220130)

[KESIMPULAN DAN REKOMENDASI 93](#_Toc220220131)

[5.1 Kesimpulan 93](#_Toc220220132)

[5.2 Implikasi Penelitian 94](#_Toc220220133)

[5.3 Keterbatasan Penelitian 94](#_Toc220220134)

[5.4 Saran untuk Penelitian Selanjutnya 95](#_Toc220220135)

[DAFTAR PUSTAKA 103](#_Toc220220136)

[**DAFTAR LAMPIRAN** 106](#_Toc220220137)

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

# BAB I

# PENDAHULUAN

## Latar Belakang

Penerapan model pembelajaran mesin (*machine learning*) pada sistem klasifikasi berbasis *computer* *vision* telah menjadi bagian integral dalam berbagai aplikasi industri, mulai dari sistem pengawasan, keselamatan kerja, hingga otomasi proses berbasis citra [1]⁠. Model-model tersebut umumnya dikembangkan dan divalidasi pada fase pelatihan menggunakan dataset yang terkontrol, dengan asumsi bahwa distribusi data pada lingkungan produksi tidak akan mengalami perubahan yang signifikan [2]⁠⁠.

Namun, dalam praktiknya, lingkungan produksi bersifat dinamis dan tidak sepenuhnya dapat diprediksi. Sistem berbasis computer vision sangat bergantung pada kualitas data visual yang diterima, sehingga perubahan kecil pada kondisi operasional—seperti variasi pencahayaan, degradasi kualitas kamera, kompresi citra, atau penurunan resolusi—dapat menyebabkan pergeseran karakteristik data input [3]⁠. Pergeseran ini berpotensi menurunkan stabilitas perilaku model meskipun arsitektur dan parameter model tetap sama [4]⁠[5]⁠.

Untuk mengantisipasi kondisi tersebut, praktik monitoring model di lingkungan produksi umumnya masih mengandalkan pendekatan berbasis satu metrik statistik. Salah satu metrik yang paling banyak digunakan adalah *Population* *Stability* *Index* (PSI) [6]⁠. PSI digunakan untuk mengukur perbedaan distribusi antara data referensi (baseline) dan data operasional, serta bersifat label-agnostic, sehingga dapat diterapkan tanpa ketergantungan pada ketersediaan *ground* *truth*. Karakteristik ini menjadikan PSI sebagai pilihan yang praktis dan relevan dalam konteks monitoring model pasca-deployment, terutama pada sistem produksi berskala besar.

Meskipun demikian, penggunaan PSI sebagai satu-satunya indikator monitoring memiliki keterbatasan yang fundamental. PSI hanya merepresentasikan perubahan distribusi data secara statistik, tanpa mempertimbangkan bagaimana perubahan tersebut memengaruhi perilaku internal model, tingkat kepercayaan prediksi, maupun performa sistem secara operasional [7]⁠[8]⁠. Dalam beberapa kondisi, nilai PSI dapat meningkat tanpa disertai penurunan kualitas prediksi yang signifikan, sementara pada kondisi lain, degradasi performa model dapat terjadi meskipun perubahan distribusi data relatif kecil.

Keterbatasan ini menunjukkan bahwa pendekatan monitoring berbasis *single*-*metric*, meskipun efektif sebagai *early* *warning*, tidak cukup untuk merepresentasikan kondisi kesehatan model secara menyeluruh. Ketergantungan pada satu metrik statistik berpotensi menghasilkan interpretasi yang parsial, ambigu, atau bahkan menyesatkan ketika digunakan sebagai dasar pengambilan keputusan operasional, seperti penentuan waktu intervensi atau rollback model [9]⁠ [9]⁠.

Oleh karena itu, diperlukan suatu pendekatan monitoring yang mampu mengintegrasikan berbagai dimensi degradasi model—tidak hanya dari sisi perubahan distribusi data, tetapi juga dari aspek kepercayaan prediksi dan performa sistem—ke dalam satu kerangka evaluasi yang terstruktur [10]⁠. Pendekatan ini diharapkan dapat memberikan representasi kondisi kesehatan model yang lebih komprehensif, objektif, dan relevan terhadap kebutuhan operasional sistem machine learning di lingkungan produksi.

.

Gambar 1.1 Ilustrasi Degradasi Model Pasca Deployment dalam Lingkungan MLOps

Sumber: Penulis, 2025

## Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diidentifikasi beberapa permasalahan utama terkait monitoring model pembelajaran mesin pada lingkungan produksi, khususnya pada sistem klasifikasi berbasis *computer* *vision*.

Pertama, praktik monitoring model yang umum diterapkan di lingkungan produksi masih didominasi oleh pendekatan berbasis satu metrik statistik, seperti PSI. Meskipun PSI efektif dalam mendeteksi pergeseran distribusi data, metrik ini hanya merepresentasikan satu dimensi degradasi, yaitu perubahan karakteristik data, tanpa memberikan gambaran mengenai dampaknya terhadap perilaku internal dan kualitas prediksi model secara keseluruhan [11]⁠.

Kedua, perubahan distribusi data yang terdeteksi oleh PSI tidak selalu berkorelasi secara langsung dengan penurunan performa atau stabilitas model. Dalam beberapa kondisi, nilai PSI dapat meningkat sebagai akibat dari perubahan data yang bersifat tidak kritis terhadap kemampuan prediksi model. Sebaliknya, degradasi kualitas prediksi dapat terjadi secara gradual tanpa diiringi perubahan PSI yang signifikan. Kondisi ini menimbulkan ambiguitas dalam interpretasi sinyal monitoring apabila hanya mengandalkan satu metrik statistik sebagai indikator utama [12]⁠.

Ketiga, pendekatan monitoring berbasis *single*-*metric* tidak mempertimbangkan aspek lain yang relevan dalam konteks operasional sistem produksi, seperti perubahan tingkat kepercayaan prediksi model, pergeseran proporsi kelas prediksi, maupun dampak degradasi terhadap performa sistem dari sisi efisiensi dan *latency*. Ketiadaan integrasi antar dimensi tersebut menyebabkan sistem monitoring tidak mampu merepresentasikan kondisi kesehatan model secara komprehensif [13]⁠.

Keempat, keputusan operasional yang diambil berdasarkan sinyal monitoring single-metric berpotensi menjadi tidak optimal. Ketergantungan pada satu metrik dapat menyebabkan keterlambatan dalam mendeteksi degradasi model yang bersifat gradual atau menghasilkan false alarm yang tidak relevan secara operasional. Hal ini menjadi tantangan tersendiri, terutama pada sistem produksi yang membutuhkan respons cepat dan tepat terhadap perubahan kondisi data.

Kelima, hingga saat ini belum terdapat pendekatan monitoring yang secara sistematis menggabungkan berbagai metrik degradasi model ke dalam satu indikator komposit yang dapat digunakan sebagai dasar pengambilan keputusan operasional. Ketiadaan kerangka evaluasi multi-dimensi ini menyulitkan proses interpretasi kondisi kesehatan model secara objektif dan konsisten.

Berdasarkan permasalahan tersebut, diperlukan suatu pendekatan monitoring model yang mampu mengintegrasikan berbagai aspek degradasi—baik dari sisi perubahan distribusi data, perilaku prediksi model, maupun performa sistem—ke dalam satu kerangka pengambilan keputusan yang terstruktur dan dapat diandalkan dalam lingkungan produksi.

Tabel 1.1 – Identifikasi Permasalahan Monitoring Model di Lingkungan Produksi

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Permasalahan** | **Dampak** | **Keterbatasan Pendekatan Existing** |
| 1 | Monitoring berbasis satu metrik | Degradasi terlambat terdeteksi | Tidak representatif |
| 2 | Tidak ada agregasi metrik | Sulit interpretasi kesehatan model | Analisis manual |
| 3 | Tidak mempertimbangkan latency | Risiko SLA production | Fokus hanya akurasi |

## Rumusan Masalah

Berdasarkan latar belakang dan identifikasi masalah yang telah diuraikan, penelitian ini berangkat dari permasalahan utama terkait efektivitas pendekatan monitoring model pembelajaran mesin di lingkungan produksi. Praktik monitoring yang umum digunakan saat ini masih didominasi oleh pendekatan berbasis satu metrik statistik, khususnya PSI, yang berfokus pada deteksi pergeseran distribusi data. Meskipun PSI relevan sebagai indikator awal terjadinya perubahan data, metrik ini memiliki keterbatasan dalam merepresentasikan kondisi kesehatan model secara menyeluruh, terutama dalam konteks sistem klasifikasi berbasis computer vision yang sangat sensitif terhadap perubahan kualitas input visual.

Permasalahan utama yang muncul adalah bagaimana keterbatasan pendekatan *single*-*metric* berbasis PSI memengaruhi kemampuan sistem monitoring dalam mendeteksi degradasi model secara tepat dan konsisten di lingkungan produksi. Ketergantungan pada satu indikator statistik berpotensi menghasilkan sinyal monitoring yang ambigu, terlambat, atau tidak selaras dengan kondisi aktual perilaku model, sehingga menyulitkan proses interpretasi dan pengambilan keputusan operasional [12]⁠.

Sejalan dengan permasalahan tersebut, muncul kebutuhan untuk mengevaluasi apakah pendekatan monitoring berbasis multi-metrik dapat memberikan representasi kondisi kesehatan model yang lebih komprehensif dibandingkan pendekatan *single*-*metric*. Integrasi berbagai metrik yang merepresentasikan dimensi degradasi yang berbeda—baik dari sisi perubahan distribusi data, perilaku prediksi model, maupun performa sistem—diharapkan mampu mengurangi bias interpretasi dan meningkatkan keandalan sistem monitoring [3]⁠.

Selain itu, permasalahan lain yang menjadi fokus penelitian ini adalah bagaimana mengintegrasikan berbagai metrik tersebut ke dalam suatu indikator komposit yang dapat digunakan sebagai dasar pengambilan keputusan operasional. Penggunaan composite health score diharapkan tidak hanya mampu mendeteksi degradasi model secara lebih dini, tetapi juga menyediakan sinyal yang lebih stabil dan konsisten untuk mendukung tindakan intervensi, seperti penyesuaian sistem atau rollback model.

Dengan demikian, penelitian ini dirumuskan untuk mengkaji secara sistematis keterbatasan pendekatan monitoring berbasis PSI sebagai single-metric serta mengeksplorasi efektivitas pendekatan multi-metrik berbasis composite health score dalam mendeteksi degradasi model pembelajaran mesin di lingkungan produksi.

## Tujuan Penelitian

Penelitian ini bertujuan untuk mengkaji dan mengevaluasi efektivitas pendekatan monitoring model pembelajaran mesin di lingkungan produksi, khususnya pada sistem klasifikasi berbasis computer vision, yang selama ini masih didominasi oleh penggunaan metrik tunggal berbasis statistik, seperti Population Stability Index (PSI). Fokus utama penelitian diarahkan untuk memahami sejauh mana pendekatan single-metric berbasis PSI mampu merepresentasikan kondisi degradasi model secara akurat dan konsisten dalam menghadapi perubahan kualitas data operasional.

Selain itu, penelitian ini bertujuan untuk mengusulkan dan mengimplementasikan pendekatan monitoring berbasis multi-metrik yang mengintegrasikan berbagai dimensi degradasi model ke dalam satu kerangka evaluasi yang terstruktur. Pendekatan ini dirancang untuk melampaui keterbatasan PSI sebagai indikator tunggal dengan mengombinasikan metrik yang merepresentasikan perubahan distribusi data, perilaku prediksi model, serta performa sistem secara operasional.

Tujuan berikutnya adalah membangun suatu composite health score berbasis multi-criteria decision model yang dapat digunakan sebagai indikator kondisi kesehatan model di lingkungan produksi. Composite health score ini diharapkan mampu memberikan sinyal monitoring yang lebih stabil, komprehensif, dan relevan secara operasional dibandingkan pendekatan single-metric.

Lebih lanjut, penelitian ini bertujuan untuk melakukan penentuan threshold pada composite health score secara empiris melalui eksperimen terkontrol pada skenario degradasi data tertentu. Penentuan threshold ini diarahkan untuk mendukung pengambilan keputusan operasional, seperti indikasi perlunya intervensi atau rollback model, berdasarkan perilaku aktual metrik monitoring, bukan semata-mata berdasarkan asumsi atau nilai heuristik.

Secara keseluruhan, penelitian ini bertujuan untuk menghasilkan kerangka monitoring model yang lebih adaptif dan aplikatif, yang tidak hanya relevan secara akademik, tetapi juga dapat diterapkan pada sistem machine learning di lingkungan produksi yang dinamis dan kompleks [14]⁠[10]⁠.

## Ruang Lingkup Penelitian

Ruang lingkup penelitian ini dibatasi untuk memastikan fokus kajian tetap terarah dan selaras dengan tujuan penelitian yang telah ditetapkan. Penelitian ini difokuskan pada monitoring degradasi model pembelajaran mesin di lingkungan produksi, khususnya pada sistem klasifikasi berbasis computer vision. Model yang diteliti merupakan model klasifikasi dengan jumlah kelas terbatas, yang dirancang dan dievaluasi pada skenario operasional yang merepresentasikan kondisi dunia nyata.

Penelitian ini hanya membahas aspek monitoring pasca-deployment dan tidak mencakup proses pelatihan ulang (*retraining*) atau adaptasi model secara online. Fokus utama diarahkan pada deteksi degradasi model berdasarkan perubahan karakteristik data input, perilaku prediksi model, dan performa sistem, tanpa melakukan modifikasi terhadap arsitektur maupun parameter model selama proses monitoring berlangsung.

Dalam konteks evaluasi metrik, penelitian ini menggunakan PSI sebagai pendekatan *baseline* *single*-*metric* untuk merepresentasikan praktik monitoring yang umum digunakan. Selanjutnya, penelitian ini mengusulkan pendekatan monitoring berbasis multi-metrik yang mengintegrasikan beberapa indikator degradasi ke dalam suatu *composite* *health* *score*. Evaluasi dan perbandingan dilakukan untuk menilai efektivitas pendekatan multi-metrik dibandingkan pendekatan *single*-*metric* berbasis PSI.

Eksperimen penentuan *threshold* pada *composite* *health* *score* dibatasi pada dua skenario degradasi data, yaitu degradasi blur dan compression. Pembatasan ini dilakukan untuk menjaga fokus eksperimen dan memastikan analisis yang mendalam terhadap perilaku metrik monitoring pada kondisi degradasi yang paling relevan terhadap sistem *computer* *vision*. Skenario degradasi lain, seperti variasi pencahayaan dan noise, digunakan sebagai bagian dari evaluasi monitoring umum, namun tidak termasuk dalam proses kalibrasi *threshold*.

Selain itu, penelitian ini tidak membahas aspek implementasi sistem monitoring secara *real*-*time* maupun integrasi dengan infrastruktur produksi secara penuh, seperti orkestrasi pipeline *Machine Learning Operations* (MLOps) atau otomatisasi *rollback*. Fokus penelitian dibatasi pada analisis metodologis dan evaluasi eksperimen berbasis simulasi lingkungan produksi untuk mendukung pengambilan keputusan operasional.

Dengan batasan ruang lingkup tersebut, penelitian ini diharapkan dapat memberikan kontribusi yang terfokus dan mendalam dalam pengembangan pendekatan monitoring model pembelajaran mesin berbasis multi-metrik, tanpa mengaburkan tujuan utama penelitian.

Tabel 1.2 Ruang Lingkup dan Batasan Penelitian

|  |  |  |
| --- | --- | --- |
| **Aspek** | **Ruang Lingkup** | **Di luar Penelitian** |
| Domain | Computer Vision | NLP, Time Series |
| Model | CNN | Transformer |
| Monitoring | Post-deployment | Training time |
| Dataset | Skenario degradasi visual | Real-world CCTV |

## Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik dari sisi akademik maupun praktis, khususnya dalam konteks monitoring model pembelajaran mesin di lingkungan produksi.

Dari sisi akademik, penelitian ini memberikan kontribusi dalam memperkaya kajian mengenai monitoring degradasi model pembelajaran mesin dengan menyoroti keterbatasan pendekatan single-metric berbasis PSI. Penelitian ini memperluas perspektif evaluasi degradasi model dengan mengusulkan pendekatan monitoring berbasis multi-metrik yang mengintegrasikan berbagai dimensi degradasi ke dalam suatu composite health score. Pendekatan ini diharapkan dapat menjadi referensi bagi penelitian selanjutnya yang membahas pemantauan kesehatan model dalam konteks MLOps dan sistem cerdas di lingkungan produksi.

Selain itu, penelitian ini memberikan kontribusi metodologis melalui penerapan *multi-criteria decision model* (MCDM) dalam konteks monitoring model pembelajaran mesin. Integrasi konsep pengambilan keputusan multi-kriteria dengan evaluasi metrik degradasi model membuka peluang pengembangan pendekatan monitoring yang lebih sistematis dan terstruktur, khususnya pada studi-studi yang menekankan aspek pengambilan keputusan berbasis data.

Dari sisi praktis, penelitian ini diharapkan dapat memberikan manfaat bagi praktisi dan pengelola sistem *machine* *learning* di lingkungan produksi dengan menyediakan kerangka monitoring yang lebih komprehensif dibandingkan pendekatan single-metric. Penggunaan *composite* *health* *score* memungkinkan sistem monitoring untuk memberikan sinyal degradasi yang lebih stabil dan relevan secara operasional, sehingga dapat mendukung pengambilan keputusan terkait intervensi atau rollback model secara lebih tepat.

Lebih lanjut, pendekatan penentuan threshold berbasis eksperimen yang diusulkan dalam penelitian ini dapat membantu praktisi dalam menetapkan batasan operasional monitoring secara objektif, tanpa bergantung sepenuhnya pada nilai heuristik atau asumsi subjektif. Dengan demikian, hasil penelitian ini diharapkan dapat menjadi acuan awal dalam perancangan sistem monitoring model yang lebih adaptif dan aplikatif pada sistem computer vision di lingkungan produksi yang dinamis.

## Hipotesis Penelitian

Berdasarkan rumusan masalah, tujuan penelitian, serta pendekatan metodologis yang digunakan, penelitian ini disusun dengan beberapa hipotesis yang akan diuji secara empiris melalui eksperimen terkontrol.

Penelitian ini mengajukan hipotesis bahwa pendekatan monitoring model berbasis *single*-*metric*, khususnya PSI, memiliki keterbatasan dalam mendeteksi degradasi model klasifikasi berbasis computer vision secara konsisten di lingkungan produksi. Keterbatasan tersebut muncul karena PSI hanya merepresentasikan perubahan distribusi data, sehingga tidak selalu mencerminkan degradasi perilaku model maupun dampaknya terhadap performa sistem secara operasional.

Selanjutnya, penelitian ini menghipotesiskan bahwa pendekatan monitoring berbasis multi-metrik, yang mengintegrasikan berbagai indikator degradasi model ke dalam suatu composite health score, mampu memberikan representasi kondisi kesehatan model yang lebih komprehensif dibandingkan pendekatan single-metric berbasis PSI. Integrasi beberapa dimensi degradasi diharapkan dapat mengurangi ambiguitas sinyal monitoring dan meningkatkan keandalan deteksi degradasi model.

Penelitian ini juga mengajukan hipotesis bahwa penggunaan composite health score memungkinkan deteksi degradasi model dilakukan secara lebih dini dan stabil, khususnya pada skenario degradasi data yang bersifat gradual. Dibandingkan dengan pendekatan single-metric, pendekatan multi-metrik diharapkan mampu memberikan indikasi degradasi yang lebih konsisten terhadap perubahan kualitas data operasional.

Selain itu, penelitian ini menghipotesiskan bahwa penentuan threshold pada composite health score yang dilakukan secara empiris melalui eksperimen terkontrol dapat menghasilkan batasan operasional yang lebih relevan dan aplikatif dibandingkan threshold yang ditetapkan secara heuristik. Threshold berbasis eksperimen diharapkan dapat mendukung pengambilan keputusan operasional, seperti indikasi perlunya intervensi atau rollback model, secara lebih objektif dan dapat dipertanggungjawabkan.

Dengan demikian, hipotesis dalam penelitian ini secara keseluruhan diarahkan untuk membuktikan bahwa pendekatan monitoring model berbasis multi-metrik dan composite health score merupakan alternatif yang lebih efektif dibandingkan pendekatan single-metric berbasis PSI dalam mendeteksi degradasi model pembelajaran mesin di lingkungan produksi.

## Sistematika Penulisan

Sistematika penulisan tesis ini disusun untuk memberikan gambaran yang terstruktur mengenai alur pembahasan penelitian dari awal hingga akhir. Penulisan tesis ini terdiri dari lima bab yang saling berkaitan sebagai berikut.

Bab I Pendahuluan menguraikan latar belakang penelitian, identifikasi masalah, rumusan masalah, tujuan penelitian, ruang lingkup penelitian, manfaat penelitian, hipotesis penelitian, serta sistematika penulisan. Bab ini memberikan konteks dan dasar pemikiran yang melatarbelakangi penelitian mengenai monitoring degradasi model pembelajaran mesin di lingkungan produksi.

Bab II Tinjauan Pustaka dan Landasan Teori membahas konsep-konsep teoretis dan penelitian terdahulu yang relevan dengan topik penelitian. Pembahasan pada bab ini meliputi degradasi model pembelajaran mesin, data drift, model monitoring, penggunaan Population Stability Index sebagai metrik monitoring, serta pendekatan multi-criteria decision model. Pada akhir bab ini, disajikan research gap yang menjadi dasar pengusulan pendekatan monitoring berbasis multi-metrik dalam penelitian ini.

Bab III Metodologi Penelitian menjelaskan secara rinci metode yang digunakan dalam penelitian, termasuk desain penelitian, objek dan lingkup penelitian, dataset dan model baseline, skenario degradasi data, mekanisme monitoring model, definisi dan normalisasi metrik, perhitungan composite health score, serta prosedur penentuan threshold berbasis eksperimen. Bab ini juga menguraikan alur eksperimen dan kriteria evaluasi yang digunakan untuk membandingkan pendekatan single-metric berbasis PSI dengan pendekatan multi-metrik.

Bab IV Hasil dan Pembahasan menyajikan hasil eksperimen yang diperoleh dari penerapan metode yang telah dirancang pada Bab III. Bab ini membahas hasil monitoring model menggunakan pendekatan single-metric dan multi-metrik, analisis waktu deteksi degradasi, evaluasi perilaku composite health score, serta pembahasan hasil penentuan threshold pada skenario degradasi yang diuji.

Bab V Penutup berisi kesimpulan yang merangkum hasil penelitian serta pembahasan implikasi penelitian. Selain itu, bab ini juga menyajikan keterbatasan penelitian dan saran untuk pengembangan penelitian selanjutnya yang berkaitan dengan monitoring dan pengelolaan model pembelajaran mesin di lingkungan produksi.

# BAB II

# TINJAUAN PUSTAKA

## 2.1 *Machine Learning* dan *Computer* *Vision* dalam Lingkungan Produksi

Machine learning (ML) telah menjadi komponen penting dalam pengembangan sistem cerdas yang mampu melakukan analisis dan pengambilan keputusan berbasis data. Dalam konteks computer vision (CV), model pembelajaran mesin digunakan untuk mengekstraksi informasi visual dari citra atau video guna melakukan tugas-tugas seperti klasifikasi objek, deteksi, dan pengenalan pola. Keberhasilan model computer vision sangat bergantung pada kemampuannya mempelajari representasi fitur visual yang relevan dari data pelatihan [15], [16], [17]⁠.

Pada fase pengembangan, model machine learning umumnya dilatih dan divalidasi menggunakan dataset yang terkontrol, dengan asumsi bahwa distribusi data pelatihan mewakili kondisi operasional yang akan dihadapi model di lingkungan produksi. Evaluasi performa pada tahap ini biasanya dilakukan menggunakan metrik seperti akurasi, precision, recall, atau F1-score, yang mengukur kemampuan model dalam memprediksi label dengan benar pada data uji yang tersedia.

Namun, karakteristik lingkungan produksi sering kali berbeda dengan lingkungan pelatihan. Sistem machine learning yang telah dideploy harus beroperasi pada data yang bersifat dinamis dan tidak sepenuhnya dapat diprediksi. Dalam sistem computer vision, kondisi operasional dapat berubah akibat variasi pencahayaan, kualitas sensor kamera, sudut pengambilan gambar, resolusi citra, serta proses kompresi data. Perubahan-perubahan ini dapat menyebabkan perbedaan karakteristik data operasional dibandingkan dengan data yang digunakan pada fase pelatihan [18]⁠.

Selain perubahan pada data input, lingkungan produksi juga memiliki keterbatasan dan tuntutan operasional yang berbeda, seperti kebutuhan akan respons yang cepat, keterbatasan sumber daya komputasi, serta kebutuhan akan sistem yang stabil dan andal dalam jangka panjang. Kondisi ini menyebabkan performa model yang tinggi pada fase pelatihan tidak selalu menjamin kinerja yang optimal setelah model diimplementasikan di lingkungan produksi [19]⁠.

A diagram of a block diagram

AI-generated content may be incorrect.

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>

Oleh karena itu, penerapan machine learning dan computer vision di lingkungan produksi tidak hanya memerlukan proses pelatihan model yang baik, tetapi juga membutuhkan mekanisme pemantauan (monitoring) yang mampu mengamati perilaku model secara berkelanjutan. Monitoring ini bertujuan untuk memastikan bahwa model tetap berfungsi sesuai dengan ekspektasi meskipun terjadi perubahan pada kondisi operasional. Pemahaman mengenai perbedaan karakteristik antara fase pelatihan dan fase produksi menjadi landasan penting dalam pengembangan pendekatan monitoring dan evaluasi degradasi model yang akan dibahas pada subbab-subsekuen [20]⁠.

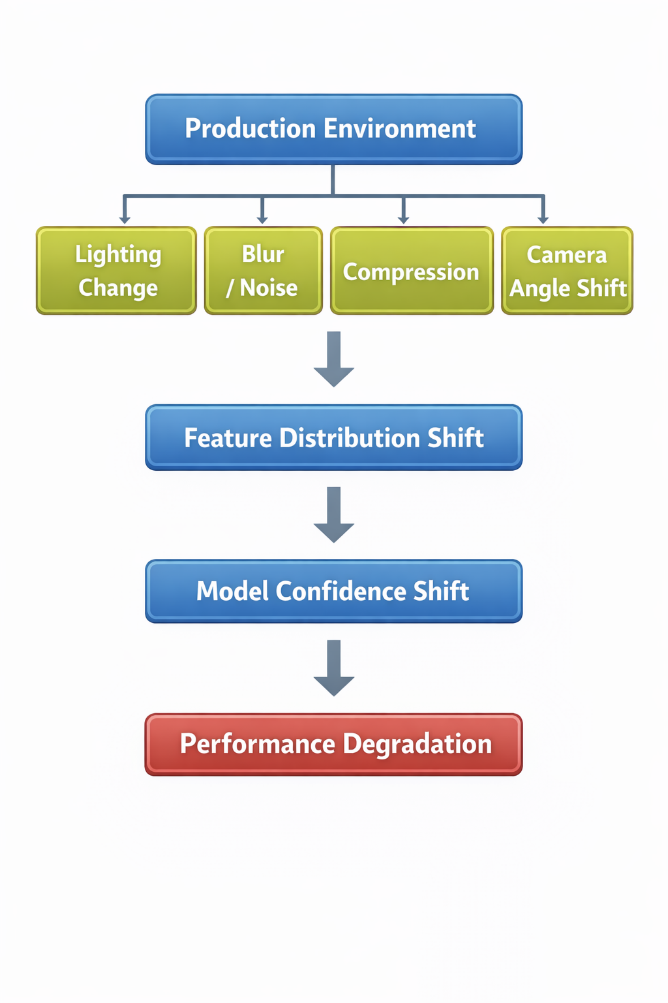
Tabel 2.1 Karakteristik Model CNN dalam Konteks Deployment Produksi

|  |  |  |
| --- | --- | --- |
| 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 Pembelajaran Mesin di Produksi

Degradasi model pembelajaran mesin merupakan fenomena di mana kinerja atau perilaku model menurun setelah model diimplementasikan di lingkungan produksi, meskipun model tersebut menunjukkan performa yang baik pada fase pelatihan dan pengujian awal. Fenomena ini sering kali muncul akibat perbedaan antara kondisi lingkungan pelatihan dan kondisi operasional yang dihadapi model secara nyata.

Dalam konteks machine learning di lingkungan produksi, degradasi model tidak selalu ditandai oleh penurunan performa secara langsung atau drastis. Pada banyak kasus, degradasi terjadi secara bertahap (gradual degradation), di mana perubahan kecil pada karakteristik data atau lingkungan operasional secara akumulatif memengaruhi stabilitas perilaku model. Hal ini menjadikan degradasi model sulit dideteksi apabila sistem monitoring hanya bergantung pada evaluasi performa yang bersifat periodik atau berbasis ground truth yang terbatas [21]⁠.



Gambar 2.2 – Ilustrasi Degradasi Model Pasca Deployment

Sumber: Pribadi, 2025

Degradasi model dapat disebabkan oleh berbagai faktor, antara lain perubahan karakteristik data input, perubahan konteks operasional, serta keterbatasan sistem pendukung model. Pada sistem computer vision, perubahan kualitas citra seperti penurunan resolusi, variasi pencahayaan, blur, dan kompresi data merupakan faktor yang umum terjadi dan berpotensi mengubah distribusi data yang diterima model. Perubahan tersebut dapat memengaruhi proses ekstraksi fitur visual dan menyebabkan pergeseran perilaku prediksi model, meskipun definisi kelas dan arsitektur model tetap sama [17], [18], [19]⁠.

Selain faktor data, degradasi model juga dapat dipengaruhi oleh faktor sistem, seperti keterbatasan sumber daya komputasi, peningkatan beban inferensi, atau perubahan latensi yang berdampak pada respons sistem secara keseluruhan. Meskipun faktor-faktor ini tidak secara langsung mengubah prediksi model, dampaknya terhadap stabilitas dan keandalan sistem tetap relevan dalam konteks operasional.

Penting untuk dicatat bahwa degradasi model tidak selalu identik dengan kegagalan model secara total. Dalam banyak situasi, model masih menghasilkan prediksi yang terlihat valid, namun tingkat kepercayaan, konsistensi, atau relevansi prediksi tersebut mengalami penurunan. Kondisi ini berpotensi menimbulkan risiko laten apabila tidak terdeteksi secara dini, terutama pada sistem yang digunakan untuk mendukung pengambilan keputusan penting.

Oleh karena itu, degradasi model pembelajaran mesin di lingkungan produksi menjadi isu yang krusial dan memerlukan perhatian khusus. Pemahaman yang tepat mengenai karakteristik dan penyebab degradasi model menjadi dasar penting dalam perancangan mekanisme monitoring yang mampu mendeteksi perubahan perilaku model secara lebih efektif. Subbab selanjutnya akan membahas konsep concept drift dan data drift sebagai salah satu mekanisme utama yang berkontribusi terhadap terjadinya degradasi model di lingkungan produksi.

Tabel 2.2 Perbandingan Jenis Drift dan Dampaknya terhadap Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Jenis Drift** | **Karakteristik Utama** | **Penyebab Umum** | **Dampak terhadap Model** | **Implikasi Monitoring** |
| **Data Drift** | Perubahan distribusi data input tanpa perubahan relasi input–output | Perubahan pencahayaan, blur, noise, resolusi, kompresi citra | Distribusi fitur berubah, confidence prediction dapat menurun | Dapat dideteksi melalui metrik statistik (PSI, KL Divergence) sebagai early warning |
| **Concept Drift** | Perubahan relasi antara input dan label target | Perubahan konteks operasional, perubahan definisi kelas, perilaku objek | Akurasi dan relevansi prediksi menurun secara langsung | Sulit dideteksi tanpa ground truth; membutuhkan indikator perilaku model |
| **Population Drift** | Perubahan proporsi kemunculan kelas atau output prediksi | Perubahan pola kejadian di lingkungan operasional | Bias prediksi dan ketidakseimbangan output kelas | Perlu monitoring distribusi output dan class ratio |
| **Quality Drift** | Penurunan kualitas data input secara bertahap | Degradasi sensor, kompresi berulang, gangguan lingkungan | Ekstraksi fitur melemah, stabilitas prediksi menurun | Tidak selalu terdeteksi oleh performa; perlu kombinasi metrik data dan output |
| **Operational Drift** | Perubahan kondisi sistem pendukung model | Beban sistem meningkat, latency bertambah | Dampak tidak langsung pada SLA dan keandalan sistem | Perlu monitoring metrik sistem seperti latency dan throughput |

## 2.3 Concept Drift dan Data Drift

Perubahan karakteristik data dan lingkungan operasional merupakan salah satu faktor utama yang memengaruhi kinerja model pembelajaran mesin di lingkungan produksi. Dalam literatur machine learning, fenomena ini umumnya dibahas melalui konsep concept drift dan data drift. Pemahaman yang jelas mengenai kedua konsep tersebut menjadi penting untuk menjelaskan sumber degradasi model serta menentukan pendekatan monitoring yang tepat [22]⁠.

2.3.1 Concept Drift

*Concept drift* merujuk pada perubahan hubungan antara data input dan label target yang dipelajari oleh model seiring waktu. Dengan kata lain, konsep yang mendasari proses pembentukan prediksi model mengalami perubahan, sehingga pola yang sebelumnya valid pada fase pelatihan tidak lagi sepenuhnya merepresentasikan kondisi operasional yang aktual.

Concept drift dapat terjadi dalam berbagai bentuk. Pada *abrupt drift*, perubahan konsep terjadi secara tiba-tiba, misalnya akibat perubahan kebijakan, regulasi, atau kondisi lingkungan yang ekstrem. Sementara itu, *gradual drift* terjadi secara perlahan, di mana konsep lama dan konsep baru muncul secara bersamaan dalam periode tertentu sebelum konsep lama sepenuhnya menghilang. Dalam konteks sistem produksi, *gradual drift* lebih umum terjadi dan cenderung lebih sulit dideteksi karena dampaknya terhadap performa model muncul secara bertahap.

Pada sistem *computer vision*, concept drift dapat muncul ketika makna semantik dari suatu kelas berubah atau ketika konteks visual yang terkait dengan suatu label mengalami pergeseran signifikan. Perubahan ini secara langsung memengaruhi akurasi dan relevansi prediksi model, sehingga concept drift sering dikaitkan dengan penurunan performa model yang dapat diukur menggunakan metrik berbasis *ground truth*. Namun, keterbatasan ketersediaan label di lingkungan produksi menjadikan deteksi concept drift sebagai tantangan tersendiri [5], [7], [22], [23]⁠.

2.3.2 Data Drift

Berbeda dengan concept drift, data drift merujuk pada perubahan distribusi data input atau output model tanpa adanya perubahan langsung pada hubungan antara input dan label target. Pada kondisi ini, definisi konsep tetap sama, tetapi karakteristik data yang diamati oleh model mengalami pergeseran dibandingkan dengan data pelatihan.

Data drift dapat terjadi akibat berbagai faktor, seperti perubahan sumber data, variasi kualitas sensor, perbedaan kondisi pencahayaan, resolusi citra, atau proses kompresi data. Dalam sistem computer vision, data drift sering kali muncul sebagai perubahan distribusi piksel, tekstur, atau fitur visual lainnya yang dapat memengaruhi proses ekstraksi fitur oleh model.

Karena data drift tidak selalu berdampak langsung pada label atau performa prediksi dalam jangka pendek, fenomena ini sering kali sulit diidentifikasi melalui metrik performa tradisional. Oleh karena itu, pendekatan monitoring berbasis statistik distribusi data menjadi relevan untuk mendeteksi perubahan karakteristik data yang berpotensi memicu degradasi model di kemudian hari [3], [9]⁠.

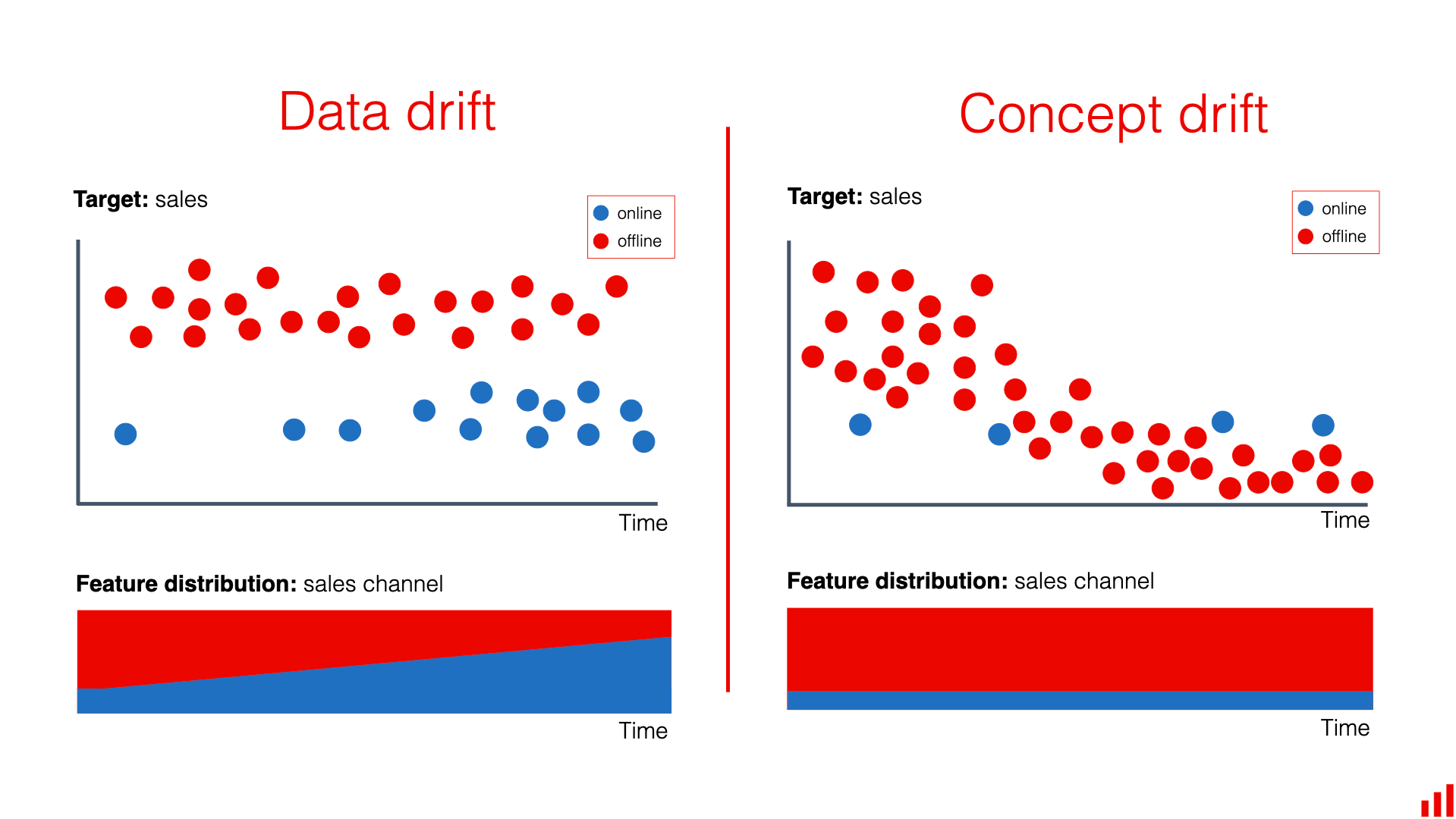
2.3.3 Hubungan Data Drift dengan Degradasi Model

Meskipun data drift dan concept drift merupakan dua konsep yang berbeda, keduanya memiliki keterkaitan yang erat dalam konteks degradasi model pembelajaran mesin. Data drift sering kali menjadi indikator awal (*early* *signal*) yang menunjukkan adanya perubahan pada lingkungan operasional, yang dalam jangka panjang dapat berkembang menjadi concept drift atau degradasi performa model yang lebih signifikan.

Namun, penting untuk dicatat bahwa tidak setiap data drift secara langsung menyebabkan penurunan performa model. Dalam beberapa kasus, model masih mampu melakukan generalisasi dengan baik meskipun terjadi perubahan distribusi data yang terdeteksi secara statistik. Sebaliknya, degradasi performa model dapat terjadi tanpa adanya indikasi data drift yang kuat, terutama ketika perubahan bersifat subtil atau terjadi pada aspek data yang tidak terwakili secara langsung oleh metrik statistik tertentu.

Kondisi ini menunjukkan bahwa data drift merupakan sinyal yang penting tetapi tidak cukup untuk merepresentasikan kondisi kesehatan model secara menyeluruh. Pendekatan monitoring yang hanya mengandalkan deteksi data drift berisiko menghasilkan interpretasi yang parsial dan kurang relevan terhadap kebutuhan pengambilan keputusan operasional. Oleh karena itu, diperlukan mekanisme monitoring yang mampu mengombinasikan informasi dari data drift dengan indikator lain yang merepresentasikan perilaku prediksi model dan performa sistem secara keseluruhan [4], [5]⁠.

Pemahaman mengenai hubungan antara data drift dan degradasi model ini menjadi landasan penting dalam pemilihan metrik monitoring serta perancangan pendekatan evaluasi yang lebih komprehensif. Subbab selanjutnya akan membahas mekanisme monitoring model pembelajaran mesin sebagai upaya untuk mendeteksi dan mengelola degradasi model di lingkungan produksi.



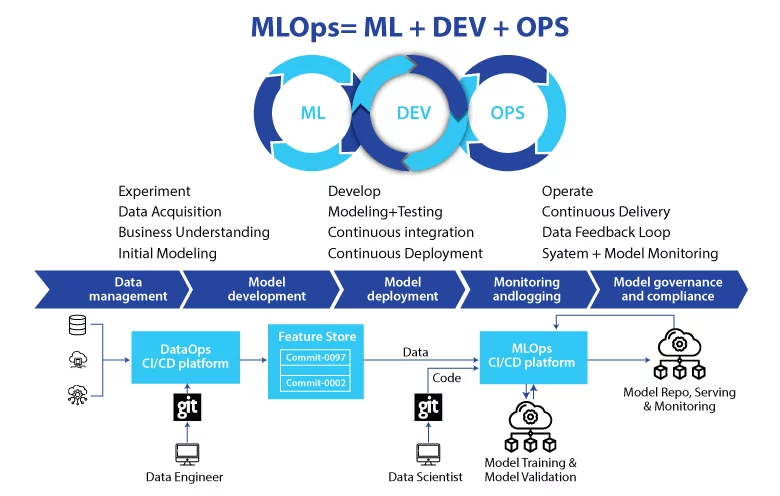
Gambar 2.3 Ilustrasi Data Drift vs Concept Drift

Sumber: <https://www.evidentlyai.com/ml-in-production/data-drift>

## 2.4 Monitoring Model Pembelajaran Mesin

Monitoring model pembelajaran mesin merupakan proses pengamatan dan evaluasi perilaku model secara berkelanjutan setelah model dideploy ke lingkungan produksi. Tujuan utama dari monitoring adalah untuk memastikan bahwa model tetap berfungsi sesuai dengan ekspektasi meskipun terjadi perubahan pada data, lingkungan operasional, maupun kondisi sistem pendukung. Dalam konteks sistem produksi, monitoring menjadi komponen penting untuk mendeteksi degradasi model secara dini dan mencegah dampak negatif terhadap kinerja sistem secara keseluruhan.

Berbeda dengan evaluasi model pada fase pelatihan, monitoring di lingkungan produksi dihadapkan pada keterbatasan ketersediaan ground truth serta dinamika data yang terus berubah. Kondisi ini menuntut pendekatan monitoring yang tidak hanya mengandalkan metrik performa tradisional, tetapi juga memanfaatkan indikator lain yang dapat merepresentasikan perubahan perilaku model secara tidak langsung. Oleh karena itu, berbagai pendekatan monitoring telah dikembangkan untuk mengamati model dari perspektif yang berbeda [16], [20]⁠.



Gambar 2.4 Alur Umum MLOps

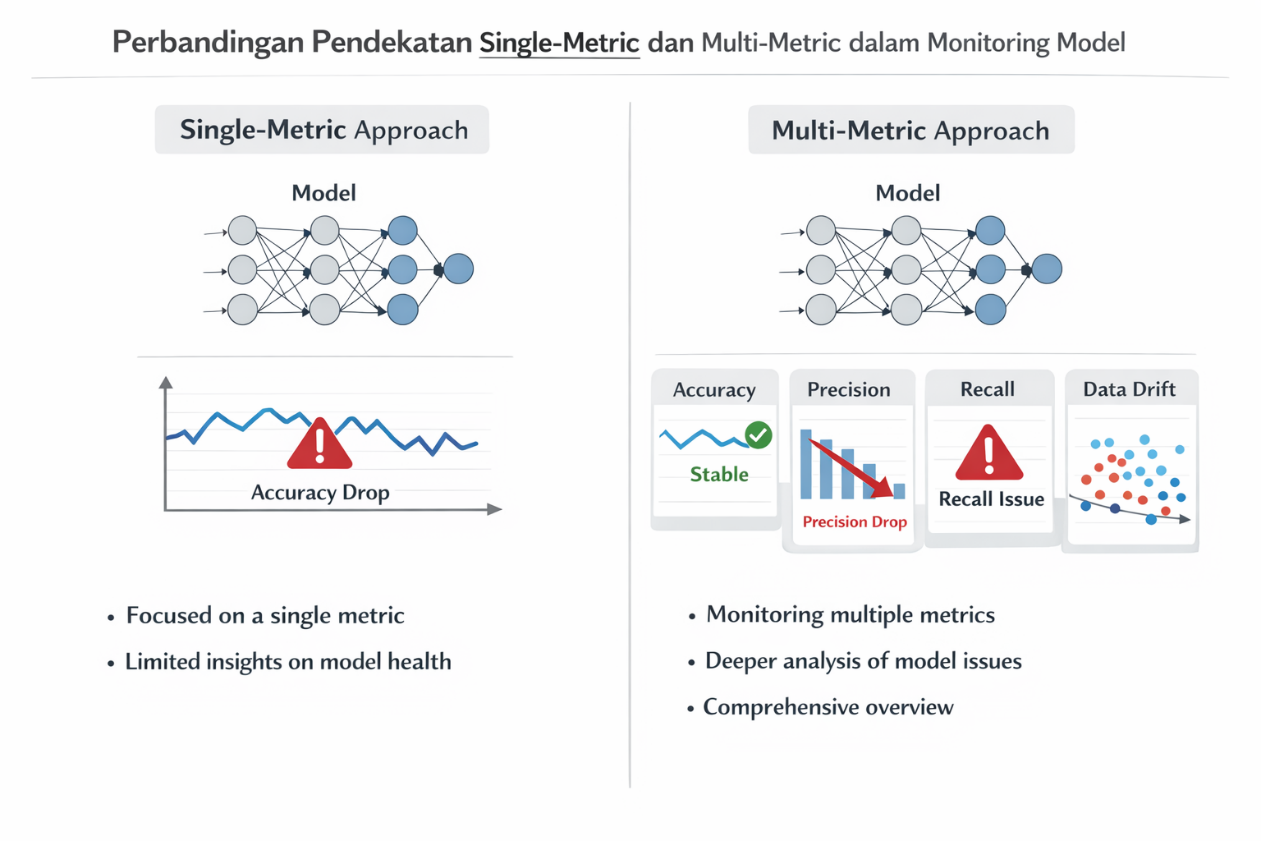
Sumber: https://www.igmguru.com/blog/machine-learning-operations-mlops-overview-definition-and-architecture

2.4.1 Monitoring Berbasis Performa

Monitoring berbasis performa berfokus pada evaluasi kinerja prediksi model menggunakan metrik seperti akurasi, precision, recall, dan F1-score. Pendekatan ini secara langsung mengukur kemampuan model dalam menghasilkan prediksi yang benar terhadap data yang tersedia. Pada fase pelatihan dan pengujian awal, metrik performa menjadi indikator utama dalam menilai kualitas model.

Namun, penerapan monitoring berbasis performa di lingkungan produksi menghadapi keterbatasan yang signifikan, terutama terkait ketersediaan ground truth. Pada banyak sistem produksi, label aktual tidak tersedia secara langsung atau hanya dapat diperoleh dengan jeda waktu tertentu melalui proses validasi manual. Akibatnya, evaluasi performa model tidak dapat dilakukan secara kontinu dan real-time.

Selain itu, monitoring berbasis performa cenderung bersifat reaktif, karena penurunan performa baru dapat terdeteksi setelah kesalahan prediksi terjadi dalam jumlah yang cukup signifikan. Pendekatan ini kurang efektif untuk mendeteksi degradasi model yang bersifat gradual, di mana perubahan perilaku model terjadi secara perlahan sebelum berdampak pada penurunan performa yang terukur [24]⁠.



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

Sumber: Penulis, 2025

2.4.2 Monitoring Berbasis Statistik

Monitoring berbasis statistik bertujuan untuk mendeteksi perubahan karakteristik data atau output model dengan menganalisis pergeseran distribusi statistik. Pendekatan ini tidak memerlukan ketersediaan ground truth, sehingga lebih sesuai untuk diterapkan di lingkungan produksi yang dinamis dan berskala besar. Monitoring statistik umumnya digunakan untuk mendeteksi data drift sebagai indikasi awal terjadinya perubahan pada lingkungan operasional.

Berbagai metrik statistik telah dikembangkan untuk tujuan ini, termasuk metrik yang mengukur perbedaan distribusi antara data referensi dan data operasional. Pendekatan ini memungkinkan sistem monitoring untuk memberikan sinyal peringatan dini (early warning) sebelum degradasi model berdampak langsung pada performa prediksi. Oleh karena itu, monitoring berbasis statistik banyak digunakan sebagai lapisan awal dalam sistem pemantauan model pembelajaran mesin.

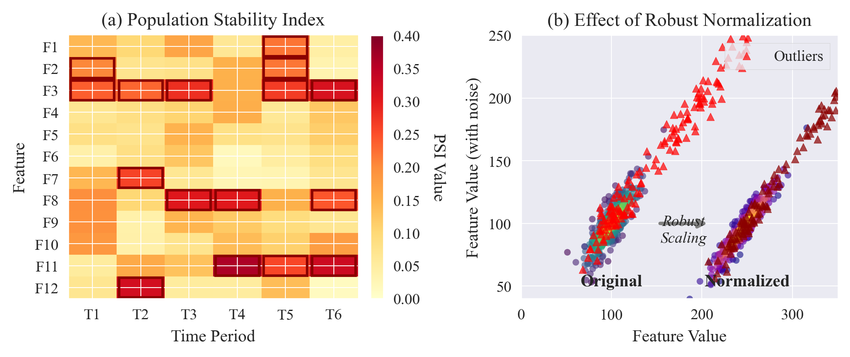
Meskipun memiliki keunggulan dalam hal ketersediaan data dan kemampuan deteksi dini, monitoring berbasis statistik juga memiliki keterbatasan. Metrik statistik umumnya hanya merepresentasikan perubahan distribusi data tanpa mempertimbangkan bagaimana perubahan tersebut memengaruhi perilaku internal model atau performa sistem secara keseluruhan. Akibatnya, sinyal yang dihasilkan dapat bersifat ambigu apabila digunakan sebagai satu-satunya dasar pengambilan keputusan operasional.

Dengan demikian, monitoring model pembelajaran mesin di lingkungan produksi memerlukan pendekatan yang mampu mengombinasikan keunggulan berbagai metode monitoring. Pemahaman mengenai perbedaan antara monitoring berbasis performa dan monitoring berbasis statistik menjadi landasan penting dalam merancang sistem monitoring yang lebih komprehensif. Subbab berikutnya akan membahas penggunaan Population Stability Index (PSI) sebagai salah satu metrik statistik yang umum digunakan dalam praktik monitoring model pembelajaran mesin.

## 2.5 Single-Metric Monitoring Berbasis Population Stability Index (PSI)

Dalam praktik monitoring model pembelajaran mesin di lingkungan produksi, pendekatan berbasis metrik statistik sering digunakan sebagai solusi yang praktis dan label-agnostic. Salah satu metrik yang paling umum diterapkan adalah Population Stability Index (PSI). PSI digunakan untuk mengukur tingkat pergeseran distribusi antara data referensi, yang umumnya berasal dari fase pelatihan atau baseline, dengan data operasional yang diamati selama model berjalan di lingkungan produksi [25]⁠.

Penggunaan PSI dalam monitoring model didorong oleh kebutuhan akan indikator yang sederhana, mudah diinterpretasikan, serta tidak bergantung pada ketersediaan ground truth. Karakteristik ini menjadikan PSI sebagai pilihan yang relevan, terutama pada sistem produksi yang berskala besar dan beroperasi secara kontinu [26]⁠.



Gambar 2.6 – Kerangka Monitoring PSI (Population Stability Index)

Sumber: https://www.researchgate.net/figure/Data-preprocessing-analysis-a-Population-Stability-Index-PSI-heatmap-showing-feature\_fig3\_396373534

2.5.1 Population Stability Index (PSI)

Population Stability Index (PSI) merupakan metrik statistik yang digunakan untuk mengukur perbedaan distribusi antara dua populasi data. Dalam konteks monitoring model pembelajaran mesin, PSI umumnya digunakan untuk membandingkan distribusi data pada kondisi baseline dengan distribusi data operasional yang diterima model setelah deployment.

PSI dihitung dengan membagi data ke dalam sejumlah bin atau interval tertentu, kemudian menghitung perbedaan proporsi data pada masing-masing bin antara dua distribusi yang dibandingkan. Nilai PSI yang rendah menunjukkan bahwa distribusi data relatif stabil, sedangkan nilai PSI yang lebih tinggi mengindikasikan terjadinya pergeseran distribusi yang semakin signifikan.

Dalam praktik industri, PSI sering digunakan sebagai indikator awal untuk mendeteksi data drift. Metrik ini banyak diterapkan karena bersifat independen terhadap model yang digunakan dan tidak memerlukan label aktual. Dengan demikian, PSI memungkinkan sistem monitoring untuk mengidentifikasi perubahan karakteristik data secara cepat tanpa menunggu hasil evaluasi performa berbasis ground truth [27]⁠.

Selain itu, PSI juga memiliki keunggulan dalam hal interpretabilitas. Nilai PSI dapat dengan mudah dipetakan ke kategori tertentu, seperti kondisi stabil, waspada, atau berisiko, sehingga memudahkan praktisi dalam memahami tingkat perubahan distribusi data yang terjadi di lingkungan produksi.

2.5.2 Keterbatasan PSI sebagai Single Metric

Meskipun PSI memiliki berbagai keunggulan sebagai metrik monitoring, penggunaan PSI sebagai satu-satunya indikator memiliki keterbatasan yang perlu diperhatikan. PSI hanya merepresentasikan perubahan distribusi data secara statistik, tanpa memberikan informasi mengenai bagaimana perubahan tersebut memengaruhi perilaku prediksi model atau performa sistem secara operasional.

Dalam beberapa kasus, nilai PSI dapat meningkat akibat perubahan data yang bersifat tidak kritis terhadap kemampuan prediksi model. Model masih dapat menghasilkan prediksi yang akurat dan konsisten meskipun terjadi pergeseran distribusi yang terdeteksi secara statistik. Sebaliknya, degradasi perilaku model dapat terjadi tanpa disertai peningkatan nilai PSI yang signifikan, terutama ketika perubahan data bersifat subtil atau terjadi pada aspek yang tidak terwakili secara langsung oleh distribusi yang diukur [22]⁠.

Selain itu, PSI tidak mempertimbangkan dimensi lain yang relevan dalam konteks kesehatan model, seperti tingkat kepercayaan prediksi, perubahan proporsi kelas hasil prediksi, atau dampak degradasi terhadap performa sistem dari sisi efisiensi dan latency. Ketergantungan pada satu metrik statistik berpotensi menghasilkan sinyal monitoring yang parsial dan kurang memadai sebagai dasar pengambilan keputusan operasional.

Keterbatasan tersebut menunjukkan bahwa meskipun PSI efektif sebagai indikator early warning terhadap perubahan distribusi data, metrik ini tidak cukup untuk merepresentasikan kondisi kesehatan model secara menyeluruh. Oleh karena itu, penggunaan PSI sebagai pendekatan single-metric perlu dilengkapi dengan metrik lain yang merepresentasikan dimensi degradasi model yang berbeda. Pendekatan ini menjadi landasan bagi pengembangan monitoring berbasis multi-metrik yang akan dibahas pada subbab berikutnya

## 2.6 Keterbatasan Pendekatan Single-Metric dalam Monitoring Model

Pendekatan monitoring model pembelajaran mesin yang mengandalkan satu metrik tunggal memiliki keterbatasan mendasar dalam merepresentasikan kompleksitas kondisi operasional di lingkungan produksi. Meskipun metrik statistik seperti Population Stability Index (PSI) efektif untuk mendeteksi perubahan distribusi data, penggunaan satu indikator saja tidak cukup untuk menggambarkan kondisi kesehatan model secara menyeluruh [11], [12]⁠.

Salah satu keterbatasan utama pendekatan single-metric adalah sifatnya yang parsial. Setiap metrik dirancang untuk mengukur aspek tertentu dari perilaku model atau data, sehingga hasil yang diperoleh hanya mencerminkan satu dimensi degradasi. Dalam konteks monitoring model, degradasi dapat muncul dari berbagai sumber, seperti perubahan karakteristik data, penurunan stabilitas prediksi, pergeseran proporsi kelas, maupun penurunan performa sistem secara operasional. Ketergantungan pada satu metrik berisiko mengabaikan dimensi degradasi lain yang sama pentingnya.

Selain itu, pendekatan single-metric rentan terhadap ambiguitas interpretasi. Nilai metrik yang menunjukkan perubahan signifikan belum tentu berkorelasi langsung dengan degradasi kinerja model yang relevan secara operasional. Sebaliknya, degradasi perilaku model dapat terjadi secara gradual tanpa memicu perubahan signifikan pada metrik tunggal yang digunakan. Kondisi ini dapat menyebabkan false alarm atau keterlambatan dalam mendeteksi degradasi model, yang pada akhirnya memengaruhi kualitas pengambilan keputusan operasional.

Keterbatasan lainnya berkaitan dengan ketidaksesuaian pendekatan single-metric terhadap kebutuhan pengambilan keputusan di lingkungan produksi. Sistem produksi tidak hanya membutuhkan informasi mengenai apakah terjadi perubahan pada data, tetapi juga memerlukan sinyal yang dapat diinterpretasikan secara langsung untuk menentukan tindakan yang tepat, seperti peningkatan pemantauan, intervensi sistem, atau rollback model. Metrik tunggal yang berdiri sendiri sering kali tidak menyediakan konteks yang cukup untuk mendukung keputusan tersebut secara objektif dan konsisten.

Lebih lanjut, pendekatan single-metric cenderung tidak adaptif terhadap variasi skenario degradasi yang berbeda. Setiap jenis degradasi dapat memengaruhi aspek perilaku model yang berbeda, sehingga sensitivitas satu metrik terhadap suatu skenario tidak selalu berlaku pada skenario lainnya. Hal ini membatasi kemampuan sistem monitoring untuk berfungsi secara andal dalam lingkungan produksi yang dinamis dan kompleks.

Berdasarkan keterbatasan tersebut, diperlukan pendekatan monitoring yang mampu mengintegrasikan berbagai metrik yang merepresentasikan dimensi degradasi model yang berbeda. Pendekatan multi-metrik diharapkan dapat memberikan representasi kondisi kesehatan model yang lebih komprehensif, mengurangi ambiguitas interpretasi, serta mendukung pengambilan keputusan operasional secara lebih efektif. Subbab berikutnya akan membahas pendekatan multi-metrik sebagai alternatif untuk mengatasi keterbatasan pendekatan single-metric dalam monitoring model pembelajaran mesin.

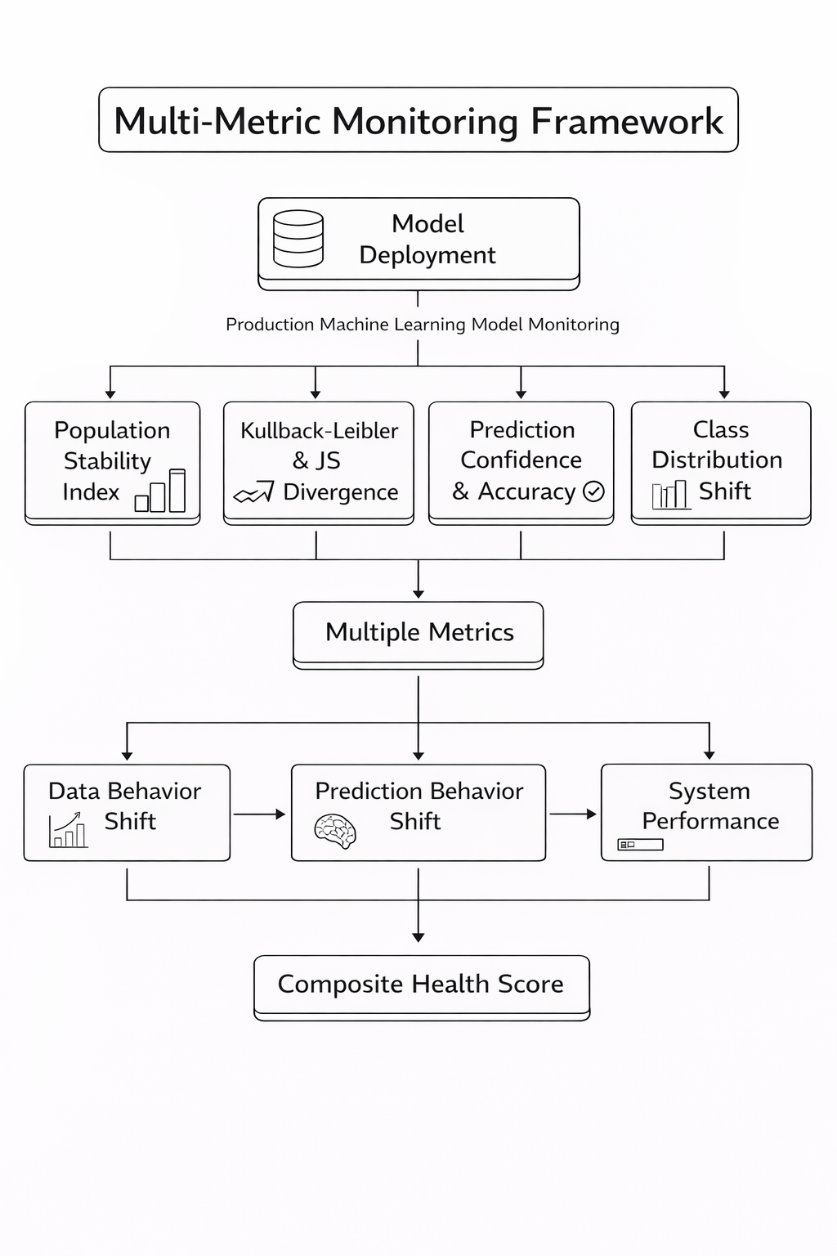
Tabel 2.3 – Perbandingan Pendekatan Single-Metric dan Multi-Metric dalam Monitoring Modelx`

|  |  |  |
| --- | --- | --- |
| **Aspek Perbandingan** | **Single-Metric Monitoring** | **Multi-Metric Monitoring** |
| **Jumlah metrik** | Satu metrik utama (mis. PSI atau confidence ratio) | Beberapa metrik yang merepresentasikan berbagai dimensi degradasi |
| **Dimensi degradasi yang diamati** | Terbatas pada satu aspek (umumnya perubahan distribusi data) | Multidimensi (distribusi data, perilaku prediksi, dan performa sistem) |
| **Ketergantungan ground truth** | Umumnya tidak memerlukan label | Umumnya tidak memerlukan label, namun dapat dikombinasikan dengan metrik performa |
| **Kemampuan deteksi dini (early warning)** | Terbatas dan bergantung pada sensitivitas satu metrik | Lebih tinggi karena menggabungkan beberapa sinyal degradasi |
| **Stabilitas sinyal monitoring** | Rentan fluktuasi dan ambigu pada kondisi tertentu | Lebih stabil karena informasi berasal dari beberapa indikator |
| **Risiko false alarm** | Relatif tinggi apabila perubahan data tidak berdampak signifikan | Lebih rendah karena perubahan dikonfirmasi oleh beberapa metrik |
| **Interpretasi kondisi model** | Parsial dan sering membutuhkan analisis tambahan | Lebih komprehensif dan representatif terhadap kondisi kesehatan model |
| **Dukungan pengambilan keputusan operasional** | Terbatas, sulit dijadikan dasar rollback | Lebih kuat, mendukung keputusan intervensi dan rollback secara objektif |
| **Kesesuaian untuk lingkungan produksi dinamis** | Kurang adaptif terhadap variasi skenario degradasi | Lebih adaptif terhadap berbagai jenis degradasi |
| **Contoh penerapan dalam penelitian ini** | PSI sebagai baseline single-metric | Composite health score berbasis multi-criteria decision model |

## 2.7 Pendekatan Multi-Metric dalam Evaluasi dan Monitoring Model

Pendekatan multi-metric dalam evaluasi dan monitoring model pembelajaran mesin muncul sebagai respons terhadap keterbatasan pendekatan single-metric yang tidak mampu merepresentasikan kompleksitas kondisi model di lingkungan produksi. Berbeda dengan pendekatan single-metric yang hanya mengandalkan satu indikator, pendekatan multi-metric mengintegrasikan beberapa metrik yang masing-masing merepresentasikan dimensi degradasi yang berbeda.

Dalam konteks monitoring model, degradasi tidak hanya dipengaruhi oleh satu faktor tunggal. Perubahan distribusi data, penurunan stabilitas prediksi, pergeseran proporsi kelas, serta degradasi performa sistem dapat terjadi secara bersamaan atau saling memengaruhi. Pendekatan multi-metric memungkinkan sistem monitoring untuk mengamati perilaku model dari berbagai perspektif, sehingga memberikan gambaran kondisi kesehatan model yang lebih komprehensif dibandingkan pendekatan berbasis satu metrik [10], [28], [29], [30], [31]⁠.



Gambar 2.7 – Kerangka Multi-Metric Monitoring

Sumber: Pribadi, 2025

Pendekatan ini juga membantu mengurangi ambiguitas interpretasi yang sering muncul pada monitoring single-metric. Dengan mempertimbangkan beberapa indikator secara bersamaan, sinyal degradasi yang dihasilkan menjadi lebih kontekstual dan relevan terhadap kondisi operasional. Misalnya, perubahan distribusi data yang terdeteksi secara statistik dapat dikaitkan dengan perubahan perilaku prediksi atau dampaknya terhadap performa sistem, sehingga interpretasi hasil monitoring menjadi lebih informatif.

Selain itu, pendekatan multi-metric lebih adaptif terhadap variasi skenario degradasi yang berbeda. Setiap jenis degradasi dapat memengaruhi aspek perilaku model yang berbeda, sehingga sensitivitas satu metrik terhadap suatu skenario tidak selalu berlaku secara universal. Dengan mengombinasikan beberapa metrik, sistem monitoring memiliki peluang yang lebih besar untuk mendeteksi degradasi secara konsisten pada berbagai kondisi operasional.

Namun demikian, penggunaan pendekatan multi-metric juga menimbulkan tantangan tersendiri, terutama terkait dengan bagaimana mengintegrasikan berbagai metrik tersebut ke dalam suatu indikator yang dapat diinterpretasikan secara praktis. Tanpa mekanisme agregasi yang terstruktur, hasil monitoring multi-metric berpotensi menjadi sulit dianalisis dan kurang efektif sebagai dasar pengambilan keputusan operasional.

Oleh karena itu, diperlukan suatu kerangka pengambilan keputusan yang mampu menggabungkan berbagai metrik degradasi model secara sistematis dan terukur. Kerangka ini diharapkan dapat mengubah informasi multi-dimensi yang dihasilkan oleh pendekatan multi-metric menjadi satu indikator komposit yang mudah diinterpretasikan dan relevan secara operasional. Pendekatan multi-criteria decision model (MCDM) menjadi salah satu solusi yang banyak digunakan untuk tujuan tersebut dan akan dibahas pada subbab selanjutnya.

## 2.8 Multi-Criteria Decision Model (MCDM)

Multi-Criteria Decision Model (MCDM) merupakan pendekatan pengambilan keputusan yang dirancang untuk menangani permasalahan kompleks yang melibatkan lebih dari satu kriteria penilaian. Dalam banyak sistem nyata, khususnya sistem teknis dan operasional, keputusan tidak dapat ditentukan hanya berdasarkan satu indikator tunggal, melainkan memerlukan pertimbangan berbagai aspek yang saling melengkapi maupun berpotensi saling bertentangan [32], [33], [34], [35]⁠.

Dalam konteks evaluasi sistem, MCDM menyediakan kerangka formal untuk menggabungkan berbagai kriteria ke dalam suatu proses penilaian yang terstruktur. Setiap kriteria merepresentasikan dimensi tertentu dari sistem yang dievaluasi, sementara bobot digunakan untuk mencerminkan tingkat kepentingan relatif dari masing-masing kriteria. Pendekatan ini memungkinkan pengambilan keputusan dilakukan secara lebih sistematis dan transparan dibandingkan pendekatan berbasis satu indikator.

A diagram of a multi-criteria decision making

AI-generated content may be incorrect.

Gambar 2.8 – Taxonomy Multi-Criteria Decision Model (MCDM)

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

2.8.1 Konsep Dasar Multi-Criteria Decision Model

Secara umum, MCDM melibatkan tiga komponen utama, yaitu alternatif, kriteria, dan bobot. Alternatif merepresentasikan objek atau kondisi yang akan dievaluasi, kriteria merupakan indikator yang digunakan untuk menilai alternatif tersebut, dan bobot mencerminkan tingkat kepentingan relatif dari setiap kriteria dalam proses pengambilan keputusan.

Proses MCDM umumnya dimulai dengan pemilihan kriteria yang relevan, diikuti dengan penentuan bobot untuk masing-masing kriteria. Selanjutnya, nilai setiap alternatif terhadap masing-masing kriteria dihitung dan digabungkan menggunakan metode agregasi tertentu untuk menghasilkan skor akhir. Skor ini kemudian digunakan sebagai dasar untuk melakukan evaluasi, perbandingan, atau pengambilan keputusan [14]⁠.

Pendekatan MCDM banyak digunakan pada berbagai bidang, seperti manajemen risiko, evaluasi kinerja sistem, perencanaan strategis, dan pengambilan keputusan berbasis data. Keunggulan utama MCDM terletak pada kemampuannya dalam mengintegrasikan berbagai dimensi penilaian ke dalam satu kerangka yang konsisten dan dapat diinterpretasikan.

2.8.2 MCDM dalam Evaluasi Sistem dan Monitoring Model

Dalam konteks evaluasi sistem pembelajaran mesin di lingkungan produksi, MCDM menjadi relevan karena kondisi kesehatan model tidak dapat direpresentasikan oleh satu metrik tunggal. Monitoring model melibatkan berbagai indikator yang mencerminkan perubahan karakteristik data, perilaku prediksi model, serta performa sistem secara operasional. Setiap indikator tersebut memberikan informasi parsial yang perlu diintegrasikan untuk memperoleh gambaran kondisi model yang utuh.

Penerapan MCDM dalam monitoring model memungkinkan berbagai metrik degradasi digabungkan secara sistematis ke dalam satu indikator komposit. Dengan pendekatan ini, sistem monitoring tidak hanya mampu mendeteksi perubahan pada masing-masing metrik secara terpisah, tetapi juga dapat mengevaluasi kondisi kesehatan model secara keseluruhan berdasarkan kombinasi berbagai kriteria yang relevan.

Selain itu, penggunaan MCDM mendukung pengambilan keputusan operasional yang lebih objektif. Dengan adanya bobot yang merepresentasikan prioritas atau kebijakan operasional tertentu, MCDM memungkinkan sistem monitoring untuk menyesuaikan sensitivitas terhadap jenis degradasi tertentu sesuai dengan kebutuhan dan konteks penggunaan sistem. Hal ini menjadikan MCDM sebagai kerangka yang fleksibel dan adaptif dalam menghadapi dinamika lingkungan produksi.

Dengan demikian, MCDM menyediakan landasan teoritis yang kuat untuk mengembangkan pendekatan monitoring model berbasis multi-metrik. Subbab berikutnya akan membahas Weighted Sum Model (WSM) sebagai salah satu metode MCDM yang dipilih dalam penelitian ini untuk mengagregasikan berbagai metrik degradasi model ke dalam suatu composite health score.

## 2.9 Weighted Sum Model (WSM) sebagai Metode Agregasi

Weighted Sum Model (WSM) merupakan salah satu metode agregasi yang paling sederhana dan umum digunakan dalam kerangka Multi-Criteria Decision Model (MCDM). Metode ini menggabungkan nilai dari beberapa kriteria dengan cara mengalikan setiap nilai kriteria dengan bobot yang telah ditentukan, kemudian menjumlahkan seluruh hasil perkalian tersebut untuk memperoleh satu skor agregat [32], [33], [34], [35], [36]⁠.

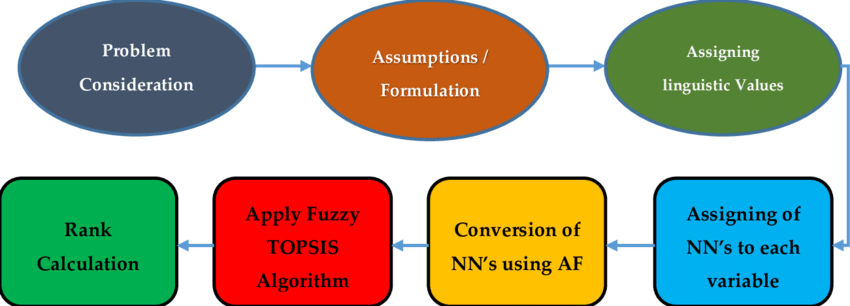
Keunggulan utama WSM terletak pada kesederhanaan dan kemudahan interpretasinya. Setiap kriteria memberikan kontribusi yang eksplisit terhadap skor akhir sesuai dengan bobot yang ditetapkan, sehingga hubungan antara nilai kriteria dan hasil evaluasi dapat dipahami dengan jelas. Karakteristik ini menjadikan WSM sesuai untuk konteks pengambilan keputusan operasional yang membutuhkan transparansi dan kemudahan penjelasan kepada pemangku kepentingan non-teknis.

Dalam konteks monitoring model pembelajaran mesin, penggunaan WSM memungkinkan berbagai metrik degradasi model yang memiliki dimensi berbeda untuk digabungkan ke dalam satu indikator komposit. Metrik-metrik tersebut dapat mencerminkan perubahan distribusi data, perilaku prediksi model, maupun performa sistem secara operasional. Dengan pendekatan WSM, setiap metrik berkontribusi terhadap composite health score sesuai dengan tingkat kepentingannya dalam kerangka monitoring yang dirancang.

Selain itu, WSM bersifat fleksibel dalam penyesuaian bobot kriteria. Bobot dapat ditentukan berdasarkan kebijakan operasional, prioritas sistem, atau pertimbangan empiris tertentu. Fleksibilitas ini penting dalam konteks lingkungan produksi, di mana kebutuhan monitoring dapat berbeda antar sistem atau berubah seiring waktu. Dengan demikian, WSM mendukung pendekatan monitoring yang adaptif tanpa memerlukan perubahan struktur metode agregasi secara fundamental.

Dibandingkan dengan metode MCDM lain yang lebih kompleks, WSM memiliki keunggulan dalam hal efisiensi komputasi dan kemudahan implementasi. Hal ini menjadikan WSM lebih sesuai untuk diterapkan dalam sistem monitoring yang berjalan secara kontinu dan memerlukan evaluasi berkala dalam waktu yang relatif singkat. Meskipun metode lain dapat menawarkan tingkat kompleksitas analisis yang lebih tinggi, kesederhanaan WSM memberikan keseimbangan yang baik antara akurasi evaluasi dan kebutuhan operasional.

Berdasarkan pertimbangan tersebut, WSM dipilih sebagai metode agregasi dalam penelitian ini untuk menggabungkan berbagai metrik degradasi model ke dalam suatu composite health score. Pemilihan WSM diharapkan dapat menghasilkan indikator kesehatan model yang mudah diinterpretasikan, relevan secara operasional, dan konsisten dengan tujuan monitoring di lingkungan produksi.



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>

## 2.10 Penelitian Terdahulu

Penelitian terdahulu yang relevan dengan topik tesis ini dapat dikelompokkan ke dalam beberapa area utama, yaitu penelitian terkait concept drift dan degradasi model, penelitian terkait monitoring model pembelajaran mesin di lingkungan produksi, serta penelitian yang menerapkan pendekatan multi-metrics dan Multi-Criteria Decision Making dalam evaluasi sistem pembelajaran mesin. Pengelompokan ini bertujuan untuk memberikan gambaran menyeluruh mengenai perkembangan penelitian serta posisi penelitian ini di antara studi-studi yang telah ada [25], [30], [31], [32]⁠.

2.10.1 Penelitian Terkait Concept Drift dan Degradasi Model

Penelitian mengenai concept drift telah banyak dilakukan untuk memahami perubahan hubungan antara data masukan dan target pada sistem pembelajaran mesin yang beroperasi secara berkelanjutan. Studi-studi awal umumnya berfokus pada pengembangan algoritma deteksi drift dengan memanfaatkan data stream berlabel, serta mengevaluasi kinerja algoritma berdasarkan metrik seperti delay deteksi, tingkat false alarm, dan akurasi prediksi. Pendekatan ini memberikan kontribusi penting dalam memahami karakteristik drift dan strategi deteksinya dalam lingkungan eksperimental.

Namun, sebagian besar penelitian pada area ini dilakukan dalam skenario terkontrol dengan asumsi ketersediaan label ground truth secara kontinu. Dalam konteks lingkungan produksi, asumsi tersebut sering kali tidak terpenuhi, sehingga pendekatan deteksi drift berbasis label menjadi sulit diterapkan secara langsung. Selain itu, fokus utama penelitian-penelitian ini adalah pada perbandingan algoritma deteksi drift, bukan pada monitoring kesehatan model secara holistik setelah deployment [22]⁠.

2.10.2 Penelitian Monitoring Model Pembelajaran Mesin

Penelitian terkait monitoring model pembelajaran mesin di lingkungan produksi umumnya menitikberatkan pada pemantauan performa prediksi dan perubahan distribusi data. Pendekatan monitoring berbasis performa menggunakan metrik seperti akurasi dan error rate untuk menilai kualitas model, sedangkan pendekatan berbasis distribusi data menggunakan metrik statistik untuk mendeteksi perubahan karakteristik data masukan.

Beberapa studi juga mulai mengeksplorasi penggunaan metrik kepercayaan model dan ketidakpastian prediksi sebagai indikator awal terjadinya degradasi model. Pendekatan ini relevan untuk lingkungan tanpa label, namun sering kali masih digunakan secara terpisah dan tidak terintegrasi dalam satu kerangka evaluasi yang terpadu. Akibatnya, interpretasi kondisi kesehatan model menjadi bergantung pada beberapa indikator terpisah yang tidak selalu memberikan sinyal yang konsisten [37]⁠.

Penelitian-penelitian ini menunjukkan adanya kesadaran terhadap keterbatasan monitoring berbasis satu metrik, namun sebagian besar masih berhenti pada penggunaan beberapa metrik secara paralel tanpa mekanisme agregasi yang sistematis.

2.10.3 Penelitian Berbasis Multi-Metrics dan MCDM

Pendekatan multi-metrics telah banyak digunakan dalam evaluasi sistem pembelajaran mesin dan algoritma data stream, khususnya untuk membandingkan kinerja beberapa alternatif berdasarkan berbagai metrik yang relevan. Dalam konteks ini, Multi-Criteria Decision Making digunakan sebagai kerangka metodologis untuk menggabungkan metrik-metrik yang heterogen ke dalam satu nilai evaluasi.

Salah satu penelitian yang menjadi rujukan utama dalam tesis ini adalah penelitian yang mengusulkan evaluasi algoritma deteksi concept drift menggunakan pendekatan Multi-Criteria Decision Making dengan metode Weighted Sum Model. Penelitian tersebut menunjukkan bahwa tidak terdapat satu metrik tunggal yang dominan untuk seluruh skenario evaluasi, dan bahwa agregasi multi-metrik menghasilkan penilaian yang lebih seimbang dan informatif dibandingkan evaluasi berbasis satu metrik. Bobot kriteria dalam penelitian tersebut ditentukan berdasarkan kebijakan evaluasi yang mencerminkan preferensi terhadap metrik tertentu [38]⁠.

Meskipun pendekatan MCDM telah terbukti efektif dalam konteks evaluasi algoritma, penerapannya masih terbatas pada skenario benchmarking dan belum secara langsung diadaptasi untuk monitoring kesehatan model pembelajaran mesin di lingkungan produksi. Selain itu, penelitian-penelitian tersebut umumnya berfokus pada evaluasi statis per dataset, bukan pada pemantauan perilaku model secara berkelanjutan dalam bentuk time-series.

2.10.4 Ringkasan Temuan Penelitian Terdahulu

Berdasarkan tinjauan terhadap penelitian terdahulu, dapat disimpulkan bahwa penelitian mengenai concept drift dan monitoring model pembelajaran mesin telah berkembang secara signifikan, namun masih terdapat keterbatasan dalam hal integrasi berbagai dimensi degradasi model ke dalam satu mekanisme monitoring yang terpadu. Pendekatan multi-metrics dan MCDM menawarkan landasan metodologis yang kuat untuk mengatasi keterbatasan tersebut, tetapi penerapannya dalam konteks monitoring kesehatan model di lingkungan produksi masih relatif terbatas.

Temuan-temuan ini menjadi dasar bagi penelitian ini untuk mengadaptasi pendekatan evaluasi multi-kriteria ke dalam mekanisme monitoring kesehatan model pembelajaran mesin, dengan fokus pada integrasi metrik degradasi yang relevan dan kebutuhan operasional sistem produksi [39]⁠.

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] | ImageNet Classification with Deep CNNs | Kinerja CNN pada klasifikasi citra | CNN (AlexNet) | CNN unggul pada tugas klasifikasi visual | Evaluasi bersifat statis (training/testing) | Menjadi dasar penggunaan CNN, belum membahas degradasi pasca-deployment |
| 2 | Howard et al., 2019 [3] | Searching for MobileNetV3 | Efisiensi CNN untuk deployment | MobileNetV3 | Model ringan dan efisien untuk inference | Tidak membahas monitoring produksi | Arsitektur model yang digunakan dalam penelitian ini |
| 3 | Lu & Xu, 2020 [5] | Understanding Model Decay in Production ML Systems | Model decay di lingkungan produksi | Analisis performa temporal | Model mengalami penurunan kinerja seiring waktu | Fokus pada performa, bukan distribusi output | Landasan konseptual degradasi model pasca-deployment |
| 4 | Gama et al., 2014 [6] | A Survey on Concept Drift Adaptation | Concept drift | Survey adaptasi model | Drift umum terjadi pada data stream | Fokus adaptasi, bukan monitoring | Menegaskan perbedaan scope (penelitian ini fokus monitoring) |
| 5 | Widmer & Kubat, 1996 [16] | Learning in the Presence of Concept Drift | Pembelajaran di bawah drift | Incremental learning | Drift memengaruhi akurasi model | Tidak membahas deployment modern | Referensi konseptual awal fenomena drift |
| 6 | Žliobaitė, 2010 [17] | Learning under Concept Drift | Definisi dan klasifikasi drift | Literature review | Drift bersifat gradual dan beragam | Tidak membahas MLOps | Landasan klasifikasi drift |
| 7 | Barandela et al., 2002 [21] | Quality Drift, Data Drift, and Concept Drift | Drift kualitas data | Analisis konseptual | Kualitas data memicu degradasi model | Tidak menyediakan mekanisme monitoring | Mendukung fokus degradasi visual berbasis distribusi |
| 8 | Carlini et al., 2019 [27] | Adversarial Examples Are Not Bugs | Sensitivitas model visual | Analisis robustness | CNN sensitif terhadap gangguan input | Fokus pada adversarial attack | Relevan untuk desain skenario degradasi visual |
| 9 | Goodfellow et al., 2016 [29] | Deep Learning | Fondasi deep learning | Teori dan praktik DL | Model sensitif terhadap perubahan distribusi data | Bukan monitoring produksi | Landasan teoretis sensitivitas CNN |
| 10 | Plumb et al., 2020 [31] | Improving Drift Detection via Probabilistic Inference | Drift berbasis probabilitas | Probabilistic inference | Deteksi drift lebih dini melalui distribusi output | Tidak dikaitkan dengan keputusan operasional | Basis monitoring probabilistik |
| 11 | Kullback & Leibler, 1951 [12] | On Information and Sufficiency | Divergensi distribusi | KL Divergence | Ukuran formal perbedaan distribusi probabilitas | Bukan konteks ML produksi | Dasar matematis KL Divergence |
| 12 | Mitchell, 1997 [11] | Machine Learning | Konsep dasar ML | Teori ML | Fondasi evaluasi dan pembelajaran mesin | Tidak membahas deployment | Landasan konseptual monitoring model |
| 13 | Sculley et al., 2015 [8] | Hidden Technical Debt in ML Systems | Risiko sistem ML produksi | Analisis sistem | Monitoring krusial untuk reliability | Tidak merinci metrik | Landasan kebutuhan monitoring dalam MLOps |
| 14 | Breck et al., 2017 [15] | The ML Test Score | Production readiness ML | Checklist MLOps | Monitoring sebagai syarat kesiapan produksi | Bersifat normatif | Mendukung kesiapan rollback model |
| 15 | Baier et al., 2021 [19] | Monitoring ML Models in Production Systems | Monitoring produksi | Survey monitoring | Monitoring ideal bersifat multidimensi | Tanpa agregasi komposit | Justifikasi pendekatan multi-metrics |
| 16 | Bhatt et al., 2021 [14] | Explainable ML in Production | Kepercayaan prediksi | Confidence-based metrics | Confidence sensitif terhadap degradasi | Tidak holistik | Baseline single-metric monitoring |
| 17 | Hwang & Yoon, 1981 [13] | Multiple Attribute Decision Making | Pengambilan keputusan multi-kriteria | Weighted Sum Model | Agregasi multi-kriteria transparan | Bukan domain ML | Dasar composite health score |
| 18 | ISO/IEC 23053, 2022 [9] | Framework for AI Systems Using ML | Kerangka AI | Standar internasional | Evaluasi lifecycle AI | Bersifat normatif | Landasan AI governance |
| 19 | ISO/IEC 5338, 2023 [10] | AI Engineering—Monitoring & Lifecycle | Monitoring sistem AI | Standar internasional | Monitoring multidimensi | Tidak teknis | Validasi tata kelola & lifecycle monitoring |

## 2.11 Posisi dan Research Gap Penelitian

Berdasarkan telaah terhadap penelitian-penelitian terdahulu yang telah diuraikan, dapat disimpulkan bahwa degradasi model pembelajaran mesin di lingkungan produksi merupakan permasalahan yang kompleks dan multi-dimensi. Berbagai pendekatan telah dikembangkan untuk mendeteksi perubahan pada data maupun performa model, namun masing-masing pendekatan masih memiliki keterbatasan dalam merepresentasikan kondisi kesehatan model secara menyeluruh dan relevan terhadap kebutuhan operasional.

Penelitian yang berfokus pada deteksi data drift dan concept drift umumnya mengandalkan metrik statistik sebagai indikator utama perubahan lingkungan operasional. Meskipun efektif sebagai sinyal awal (early warning), pendekatan ini cenderung hanya merepresentasikan satu dimensi degradasi, yaitu perubahan distribusi data, tanpa mempertimbangkan dampaknya terhadap perilaku prediksi model maupun performa sistem secara operasional. Akibatnya, hasil deteksi drift sering kali sulit diinterpretasikan secara langsung sebagai dasar pengambilan keputusan operasional.

Di sisi lain, pendekatan monitoring berbasis performa prediksi memberikan gambaran langsung mengenai kualitas model, namun menghadapi keterbatasan serius di lingkungan produksi akibat ketersediaan ground truth yang terbatas atau tertunda. Pendekatan ini bersifat reaktif dan kurang efektif dalam mendeteksi degradasi model yang terjadi secara gradual sebelum berdampak signifikan terhadap performa yang terukur.

Beberapa penelitian mulai mengarah pada penggunaan pendekatan multi-metrik untuk mengevaluasi perilaku model secara lebih komprehensif. Namun, pada banyak kasus, integrasi berbagai metrik tersebut masih bersifat deskriptif dan belum dilengkapi dengan kerangka agregasi yang terstruktur untuk mendukung pengambilan keputusan operasional. Selain itu, penerapan pendekatan Multi-Criteria Decision Model (MCDM) dalam konteks monitoring kesehatan model pembelajaran mesin di lingkungan produksi masih relatif terbatas dan umumnya difokuskan pada evaluasi atau perbandingan algoritma, bukan pada monitoring satu model yang berjalan secara operasional [10], [14], [39]⁠.

Berdasarkan kondisi tersebut, terdapat research gap dalam pengembangan pendekatan monitoring model pembelajaran mesin yang mampu mengintegrasikan berbagai dimensi degradasi—meliputi perubahan distribusi data, perilaku prediksi model, dan performa sistem—ke dalam satu indikator yang terstruktur dan mudah diinterpretasikan. Selain itu, masih terbatas penelitian yang membahas penentuan threshold monitoring secara empiris untuk mendukung pengambilan keputusan operasional, seperti indikasi intervensi atau rollback model.

Penelitian ini memposisikan diri untuk mengisi research gap tersebut dengan mengusulkan pendekatan monitoring model berbasis multi-metrik menggunakan composite health score yang dibangun melalui kerangka Multi-Criteria Decision Model. Berbeda dengan penelitian terdahulu, penelitian ini menggunakan Population Stability Index sebagai baseline single-metric yang merepresentasikan praktik monitoring umum, kemudian membandingkannya dengan pendekatan multi-metrik yang mengintegrasikan berbagai indikator degradasi model. Selain itu, penelitian ini menekankan penentuan threshold composite health score secara empiris melalui eksperimen terkontrol pada skenario degradasi data tertentu, sehingga hasil monitoring yang diperoleh lebih relevan dan aplikatif dalam konteks operasional.

Dengan demikian, posisi penelitian ini berada pada irisan antara monitoring degradasi model pembelajaran mesin dan pengambilan keputusan berbasis multi-kriteria di lingkungan produksi. Pendekatan yang diusulkan diharapkan dapat memberikan kontribusi metodologis dan praktis dalam pengembangan sistem monitoring model yang lebih komprehensif, objektif, dan dapat diandalkan untuk mendukung pengelolaan model pembelajaran mesin secara berkelanjutan.

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 |

# BAB III

# METODOLOGI PENELITIAN

## 3.1 Desain dan Pendekatan Penelitian

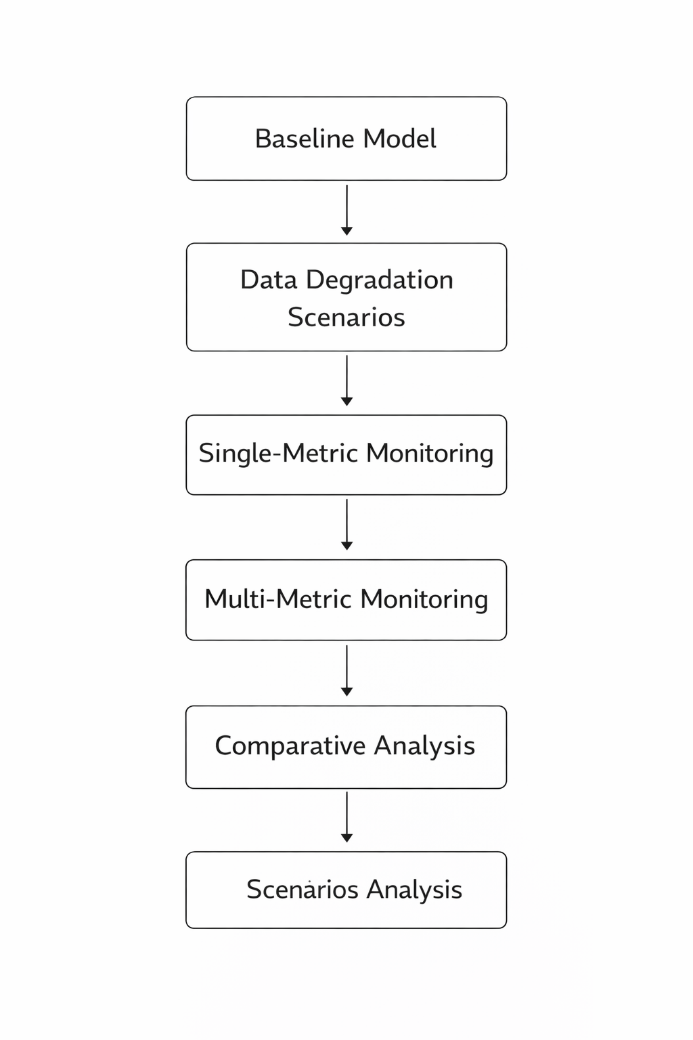
Penelitian ini menggunakan desain penelitian eksperimental kuantitatif dengan pendekatan evaluasi komparatif untuk mengkaji efektivitas metode monitoring degradasi model pembelajaran mesin di lingkungan produksi. Pendekatan ini dipilih karena penelitian berfokus pada pengamatan perilaku model secara terukur melalui eksperimen terkontrol, serta membandingkan hasil monitoring yang diperoleh dari dua pendekatan yang berbeda.

Fokus utama penelitian diarahkan pada monitoring model pasca-deployment, di mana model telah dilatih dan dievaluasi sebelumnya, kemudian dijalankan pada kondisi operasional yang mensimulasikan lingkungan produksi. Penelitian ini tidak bertujuan untuk mengembangkan atau membandingkan arsitektur model pembelajaran mesin, melainkan untuk mengevaluasi pendekatan monitoring yang digunakan dalam mendeteksi degradasi perilaku model akibat perubahan karakteristik data operasional.

Pendekatan komparatif dalam penelitian ini dilakukan dengan membandingkan monitoring berbasis single-metric dan monitoring berbasis multi-metrik. Pendekatan single-metric menggunakan Population Stability Index (PSI) sebagai baseline, karena metrik ini merepresentasikan praktik monitoring yang umum digunakan untuk mendeteksi pergeseran distribusi data di lingkungan produksi. Sementara itu, pendekatan multi-metrik mengintegrasikan beberapa indikator degradasi model ke dalam suatu composite health score menggunakan kerangka Multi-Criteria Decision Model dengan metode agregasi Weighted Sum Model.

Eksperimen dirancang untuk mengamati perbedaan kemampuan kedua pendekatan tersebut dalam mendeteksi degradasi model, khususnya pada kondisi degradasi data yang bersifat gradual. Evaluasi dilakukan berdasarkan perilaku metrik monitoring yang dihasilkan pada setiap batch data operasional, dengan menitikberatkan pada stabilitas sinyal, waktu deteksi degradasi, dan relevansinya terhadap pengambilan keputusan operasional.

Dengan desain penelitian ini, diharapkan dapat diperoleh pemahaman yang objektif mengenai keterbatasan pendekatan monitoring berbasis single-metric serta keunggulan pendekatan multi-metrik dalam merepresentasikan kondisi kesehatan model pembelajaran mesin di lingkungan produksi. Pendekatan metodologis yang digunakan pada penelitian ini selanjutnya diuraikan secara rinci pada subbab-subsekuen dalam Bab III.



Gambar 3.1 – Desain Penelitian Eksperimental  
Sumber: Pribadi, 2025

## 3.2 Objek dan Lingkup Penelitian

Objek penelitian dalam penelitian ini adalah satu model pembelajaran mesin berbasis computer vision yang digunakan untuk tugas klasifikasi citra. Model tersebut telah melalui proses pelatihan dan evaluasi awal pada fase pengembangan, kemudian digunakan sebagai objek pengamatan dalam konteks monitoring pasca-deployment. Fokus penelitian tidak diarahkan pada pengembangan atau perbandingan arsitektur model, melainkan pada evaluasi pendekatan monitoring yang diterapkan terhadap model yang telah dideploy.

Lingkup penelitian dibatasi pada monitoring perilaku model di lingkungan produksi yang disimulasikan, dengan asumsi bahwa model dijalankan secara kontinu dan menerima data operasional yang karakteristiknya dapat berubah seiring waktu. Monitoring dilakukan pada level batch data, di mana setiap batch merepresentasikan satu periode operasional tertentu. Pendekatan batch-based ini dipilih untuk mencerminkan praktik monitoring yang umum diterapkan di lingkungan produksi, terutama pada sistem yang tidak melakukan evaluasi per-inference secara real-time.

Penelitian ini tidak mencakup proses retraining, fine-tuning, atau adaptasi model secara online. Arsitektur model dan parameter yang digunakan diasumsikan bersifat tetap selama proses eksperimen berlangsung. Dengan demikian, perubahan perilaku model yang diamati dalam penelitian ini sepenuhnya dikaitkan dengan perubahan karakteristik data operasional dan kondisi sistem, bukan akibat modifikasi internal pada model.

Dalam konteks evaluasi monitoring, penelitian ini membandingkan dua pendekatan utama, yaitu pendekatan single-metric dan pendekatan multi-metrik. Pendekatan single-metric menggunakan Population Stability Index (PSI) sebagai baseline untuk merepresentasikan praktik monitoring yang umum digunakan dalam mendeteksi pergeseran distribusi data. Pendekatan multi-metrik mengintegrasikan beberapa indikator degradasi model yang merepresentasikan dimensi perubahan data, perilaku prediksi model, dan performa sistem ke dalam suatu composite health score.

Lingkup eksperimen dalam penelitian ini dibatasi pada skenario degradasi data yang bersifat gradual dan realistis, khususnya degradasi blur dan compression, yang sering dijumpai pada sistem computer vision di lingkungan produksi. Skenario degradasi lain, seperti variasi pencahayaan dan noise, dapat digunakan sebagai observasi tambahan, namun tidak menjadi fokus utama dalam penentuan threshold monitoring.

Selain itu, penelitian ini tidak membahas implementasi sistem monitoring secara end-to-end dalam infrastruktur produksi yang sesungguhnya, seperti integrasi dengan pipeline MLOps, orkestrasi deployment, atau otomatisasi tindakan rollback. Fokus penelitian dibatasi pada aspek metodologis dan evaluasi eksperimen untuk mengkaji efektivitas pendekatan monitoring dalam mendeteksi degradasi model.

Dengan batasan objek dan lingkup tersebut, penelitian ini diharapkan dapat memberikan hasil yang terfokus dan dapat dipertanggungjawabkan secara metodologis, serta relevan untuk digunakan sebagai dasar pengembangan sistem monitoring model pembelajaran mesin di lingkungan produksi.

Tabel 3.1 – Objek dan Lingkup Penelitian

|  |  |
| --- | --- |
| **Aspek** | **Deskripsi** |
| **Objek Penelitian** | Model pembelajaran mesin berbasis *computer vision* untuk tugas klasifikasi citra dua kelas (binary classification) |
| **Jenis Model** | Convolutional Neural Network (CNN) dengan arsitektur MobileNetV3 |
| **Status Model** | Model telah dilatih dan dibekukan (*frozen model*), tanpa proses retraining atau adaptasi selama eksperimen |
| **Lingkungan Evaluasi** | Lingkungan produksi tersimulasikan (*production-like environment*) |
| **Pendekatan Monitoring** | Monitoring pasca-deployment berbasis batch (*batch-based monitoring*) |
| **Pendekatan Pembanding** | Single-metric monitoring (Population Stability Index / PSI) dan multi-metric monitoring (composite health score) |
| **Dimensi Monitoring** | Perubahan distribusi data, stabilitas prediksi, pergeseran proporsi kelas, dan performa sistem (latency) |
| **Skenario Data** | Data baseline (tanpa degradasi) dan data terdegradasi |
| **Jenis Degradasi Data** | Blur, compression (JPEG), lighting variation, dan Gaussian noise |
| **Fokus Analisis** | Deteksi degradasi model dan perbandingan efektivitas pendekatan single-metric dan multi-metric |
| **Aspek yang Tidak Diteliti** | Retraining model, adaptasi online learning, deployment real-time, dan otomasi rollback produksi |
| **Unit Analisis** | Batch data operasional pada setiap skenario degradasi |
| **Tujuan Lingkup** | Mengevaluasi kemampuan monitoring dalam mendeteksi degradasi model secara dini dan konsisten |

## 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 |

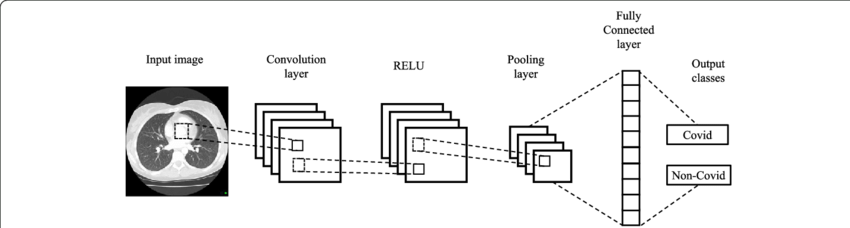
3.3.1 Arsitektur Model Klasifikasi

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 realistis. 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.



Gambar 3.3 – Arsitektur CNN Baseline

Sumber: <https://www.researchgate.net/figure/Basic-CNN-architecture-for-image-classification_fig2_351361078>

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 Alur Metode Penelitian

Alur metodologi penelitian ini dirancang untuk menggambarkan secara sistematis tahapan eksperimen yang dilakukan dalam mengevaluasi pendekatan monitoring degradasi model pembelajaran mesin di lingkungan produksi. Alur ini mencerminkan proses end-to-end mulai dari pembentukan kondisi baseline hingga evaluasi hasil monitoring dan penentuan threshold operasional.

Secara umum, metodologi penelitian diawali dengan penyiapan model baseline yang telah dilatih menggunakan dataset penelitian. Model ini digunakan sebagai referensi tetap selama seluruh eksperimen, sehingga setiap perubahan perilaku yang diamati dapat dikaitkan secara langsung dengan perubahan karakteristik data operasional, bukan akibat modifikasi internal model.

Tahapan berikutnya adalah simulasi degradasi data yang merepresentasikan perubahan kualitas data input pada lingkungan produksi. Data operasional disusun dalam bentuk batch, di mana setiap batch mencerminkan satu periode operasional tertentu. Pada tahap ini, degradasi data diterapkan secara bertahap (gradual) untuk mengamati perubahan perilaku model dari waktu ke waktu.

Setelah data operasional disiapkan, model baseline digunakan untuk melakukan proses inferensi pada setiap batch data. Output probabilitas yang dihasilkan dari proses inferensi ini kemudian dikumpulkan dan digunakan sebagai dasar perhitungan berbagai metrik monitoring. Proses ini dilakukan secara konsisten untuk setiap batch agar perubahan metrik dapat diamati secara temporal.

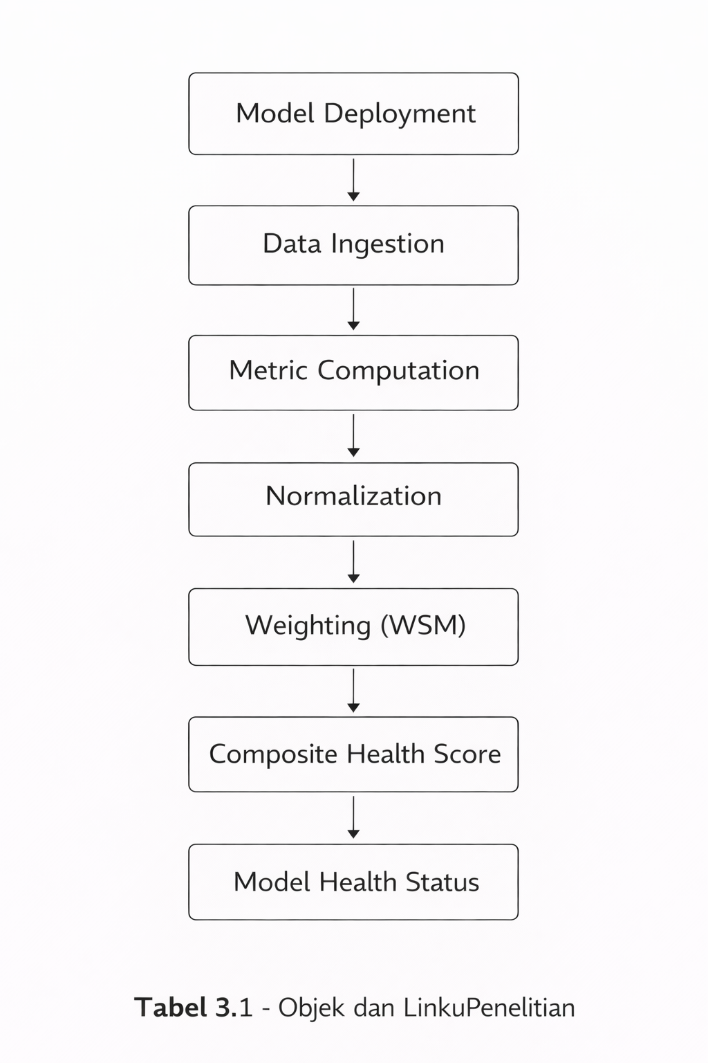
Metrik monitoring yang dihitung mencakup metrik statistik dan metrik berbasis perilaku prediksi model. Untuk pendekatan single-metric, Population Stability Index (PSI) digunakan untuk mengukur pergeseran distribusi antara data baseline dan data operasional pada setiap batch. Sementara itu, pada pendekatan multi-metrik, beberapa indikator degradasi model dihitung secara paralel untuk merepresentasikan berbagai dimensi perubahan yang terjadi.

Nilai metrik yang diperoleh selanjutnya melalui proses normalisasi untuk memastikan bahwa seluruh metrik berada pada skala yang sebanding dan dapat digabungkan secara adil. Proses normalisasi dilakukan berdasarkan nilai referensi baseline dan rentang nilai yang diperoleh dari eksperimen awal. Tahap ini menjadi prasyarat penting sebelum dilakukan agregasi metrik dalam pendekatan multi-metrik.

Pada pendekatan multi-metrik, metrik yang telah dinormalisasi digabungkan menggunakan kerangka Multi-Criteria Decision Model dengan metode agregasi Weighted Sum Model. Hasil agregasi ini menghasilkan suatu composite health score yang merepresentasikan kondisi kesehatan model pada setiap batch data operasional.

Tahapan terakhir dalam alur metodologi penelitian adalah evaluasi hasil monitoring dan penentuan threshold operasional. Composite health score yang dihasilkan dianalisis untuk mengidentifikasi pola perubahan dan titik transisi yang mengindikasikan terjadinya degradasi model. Proses ini dilakukan melalui eksperimen terkontrol pada skenario degradasi tertentu untuk memperoleh threshold yang diturunkan secara empiris.

Alur metodologi penelitian ini menjadi dasar bagi seluruh eksperimen dan analisis yang dilakukan pada penelitian ini. Dengan alur yang terstruktur dan konsisten, penelitian ini memastikan bahwa evaluasi pendekatan monitoring dilakukan secara sistematis dan dapat dipertanggungjawabkan secara metodologis.

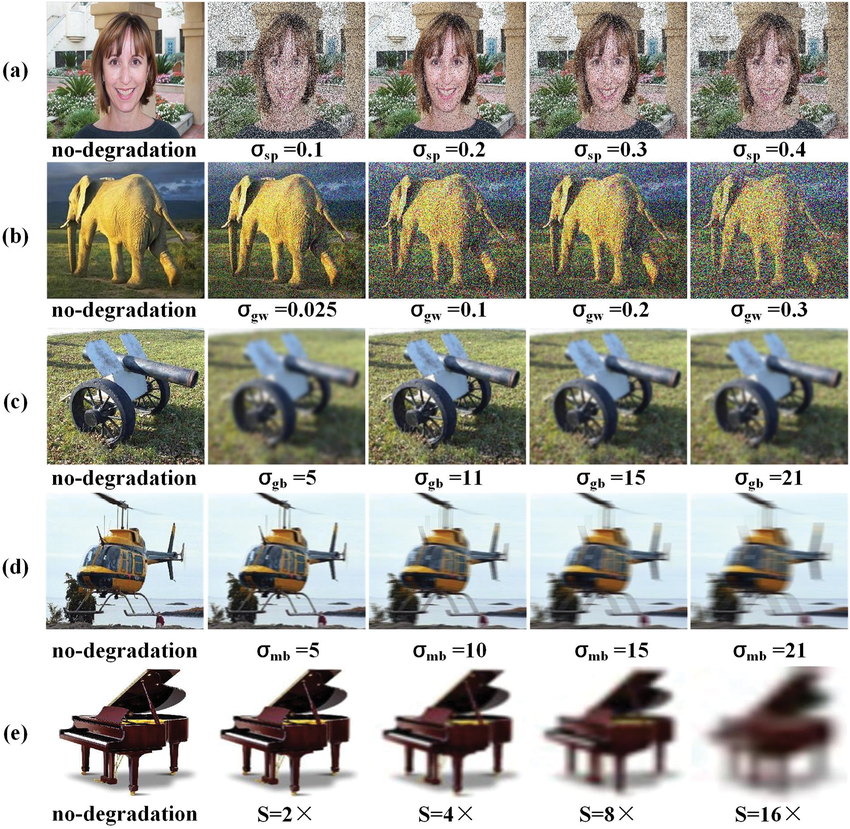


Gambar 3.4 – Alur Metodologi Monitoring Model

Sumber: Pribadi, 2025

## 3.5 Skenario Degradasi Data

Skenario degradasi data dalam penelitian ini dirancang untuk merepresentasikan berbagai kondisi operasional yang umum ditemui pada sistem computer vision di lingkungan produksi. Tujuan utama perancangan skenario ini adalah untuk mengevaluasi kemampuan mekanisme monitoring—baik berbasis single-metric maupun multi-metrik—dalam mendeteksi perubahan perilaku model pada berbagai kondisi kualitas data input.



Gambar 3.5 – Contoh Skenario Degradasi Visual

Sumber: <https://www.researchgate.net/figure/Examples-of-degraded-images-with-different-degradation-levels-a-to-f-are_fig2_377035428>

Secara keseluruhan, penelitian ini mengevaluasi mekanisme monitoring pada lima kondisi skenario, yaitu satu kondisi baseline tanpa degradasi dan empat kondisi degradasi data. Pendekatan ini memungkinkan evaluasi monitoring dilakukan secara komprehensif pada berbagai jenis gangguan kualitas data yang berbeda karakteristik dan dampaknya terhadap model.

Degradasi data diterapkan secara terkontrol dan bertahap (gradual degradation), sehingga perubahan perilaku model dan sinyal monitoring dapat diamati secara temporal. Seluruh skenario degradasi dirancang tanpa mengubah label data, sehingga perubahan yang teramati mencerminkan degradasi kualitas data input, bukan perubahan definisi kelas.

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 |

3.5.2 Table Parameter Konfigurasi Transformasi Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Jenis Degradasi** | **Parameter** | **Nilai** | **Tujuan** |
| Blur | Kernel size | 2.0 | Simulasi blur ringan–sedang |
| Compression | JPEG quality | 25 | Simulasi penurunan kualitas |
| Lighting | Brightness factor | 0.1 | Simulasi pencahayaan |
| Noise | σ (Gaussian) | 15.0 | Simulasi noise sensor |

3.5.1 Prinsip Perancangan Skenario Degradasi

Perancangan skenario degradasi data dalam penelitian ini didasarkan pada beberapa prinsip utama. Pertama, skenario degradasi harus relevan dengan kondisi operasional sistem computer vision di lingkungan produksi. Kedua, degradasi diterapkan secara bertahap untuk memungkinkan pengamatan perubahan sinyal monitoring dari kondisi normal menuju kondisi terdegradasi.

Ketiga, skenario degradasi dirancang agar bersifat label-invariant, sehingga perubahan perilaku model dapat dikaitkan secara langsung dengan kualitas data input. Keempat, seluruh skenario dapat direproduksi secara konsisten pada seluruh eksperimen untuk menjaga keterbandingan hasil monitoring antar kondisi.

3.5.2 Skenario Baseline (Tanpa Degradasi)

Skenario baseline merepresentasikan kondisi operasional normal di mana data input memiliki kualitas yang sebanding dengan data pelatihan. Pada skenario ini, tidak diterapkan degradasi apa pun terhadap data input. Skenario baseline digunakan sebagai acuan untuk mengamati perilaku model dalam kondisi stabil serta sebagai pembanding terhadap seluruh skenario degradasi data.

Hasil monitoring pada skenario baseline diharapkan menunjukkan sinyal yang relatif stabil dan merepresentasikan kondisi kesehatan model yang baik. Skenario ini menjadi referensi penting dalam analisis perubahan sinyal monitoring pada skenario degradasi berikutnya

3.5.3 Skenario Degradasi Blur

Skenario degradasi blur dirancang untuk merepresentasikan penurunan ketajaman citra yang umum terjadi pada sistem computer vision di lingkungan produksi, seperti akibat pergerakan kamera, getaran perangkat, atau fokus lensa yang tidak optimal. Degradasi blur diterapkan secara bertahap dengan meningkatkan tingkat blur pada setiap batch data operasional.

Perubahan tingkat blur menyebabkan berkurangnya detail visual yang diekstraksi oleh model, sehingga berpotensi memengaruhi stabilitas prediksi dan distribusi fitur visual. Skenario ini digunakan untuk mengevaluasi sensitivitas mekanisme monitoring terhadap degradasi yang berdampak langsung pada kualitas fitur citra.

3.5.4 Skenario Degradasi Compression (JPEG)

Skenario degradasi compression dirancang untuk mensimulasikan penurunan kualitas citra akibat proses kompresi data, seperti kompresi JPEG, yang sering digunakan dalam sistem produksi untuk menghemat bandwidth atau penyimpanan. Tingkat kompresi ditingkatkan secara bertahap pada setiap batch data operasional.

Degradasi akibat kompresi dapat menyebabkan hilangnya informasi visual penting dan perubahan distribusi nilai piksel, yang pada gilirannya dapat memengaruhi perilaku prediksi model. Skenario ini merepresentasikan kondisi operasional yang realistis, khususnya pada sistem computer vision yang berjalan pada infrastruktur dengan keterbatasan sumber daya.

3.5.5 Skenario Degradasi Lighting

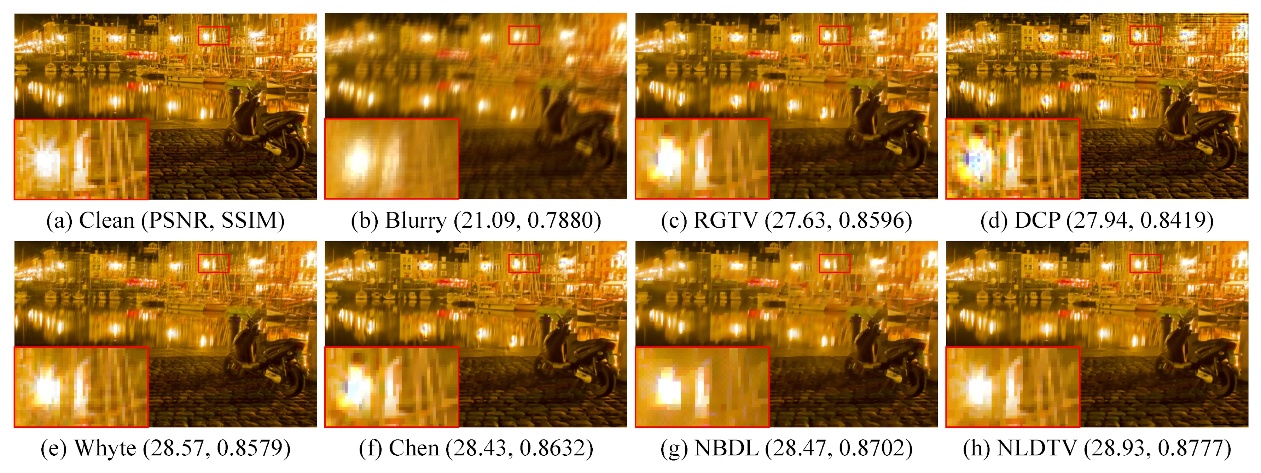
Skenario degradasi *lighting* dirancang untuk merepresentasikan perubahan kondisi pencahayaan yang sering terjadi pada lingkungan operasional nyata, seperti variasi intensitas cahaya, bayangan, atau kondisi pencahayaan rendah. Perubahan pencahayaan diterapkan secara bertahap untuk mengamati dampaknya terhadap distribusi citra dan respons model.

Perubahan kondisi pencahayaan dapat memengaruhi kontras dan representasi visual objek dalam citra, sehingga berpotensi menyebabkan pergeseran distribusi data yang tidak selalu terdeteksi secara langsung oleh metrik berbasis performa. Skenario ini digunakan untuk mengevaluasi kemampuan monitoring dalam mendeteksi degradasi yang bersifat kontekstual.

3.5.6 Skenario Degradasi Gaussian Noise

Skenario degradasi *Gaussian noise* dirancang untuk mensimulasikan gangguan acak pada citra yang dapat muncul akibat keterbatasan sensor, interferensi sinyal, atau proses akuisisi data yang tidak ideal. *Noise* ditambahkan secara bertahap pada data input dengan tingkat intensitas yang meningkat pada setiap batch data operasional.

Penambahan *noise* menyebabkan distorsi acak pada citra yang dapat memengaruhi stabilitas prediksi model dan distribusi fitur visual. Skenario ini digunakan untuk mengevaluasi robustnes mekanisme monitoring terhadap degradasi yang bersifat stokastik dan tidak terstruktur.



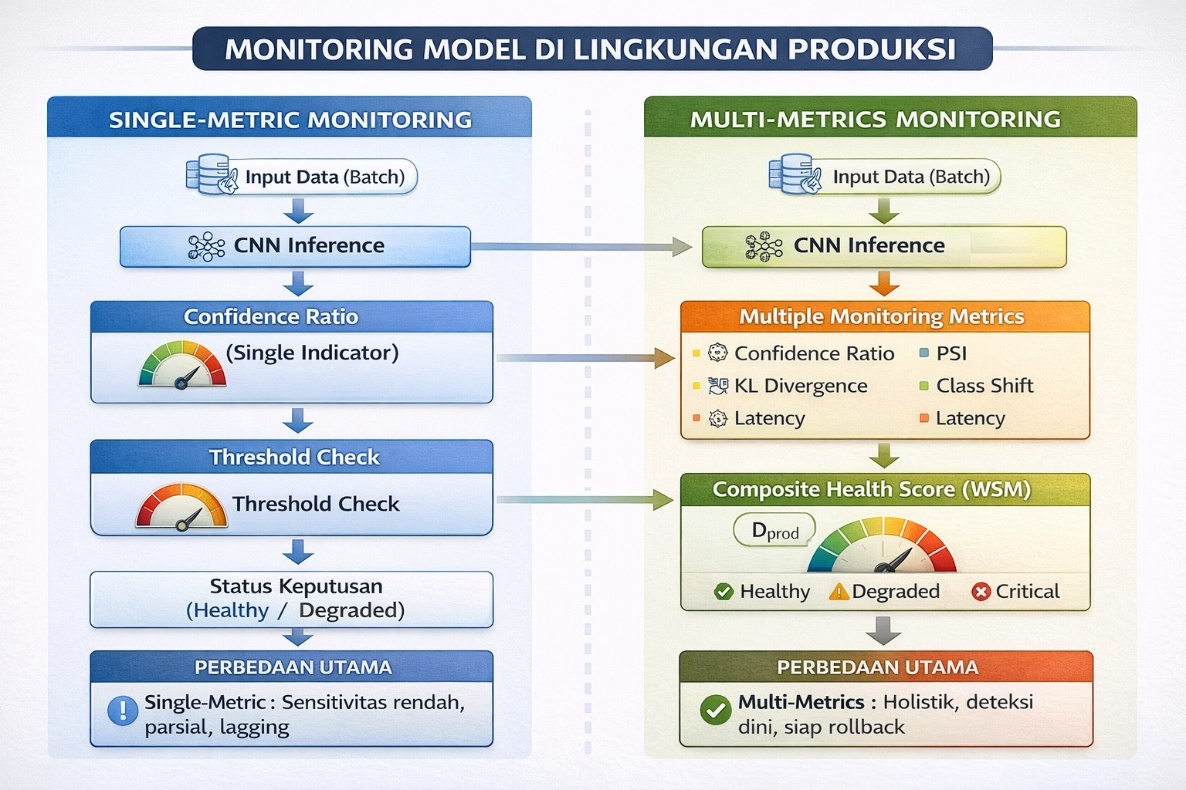
Gambar 3.4 Contoh Skenario Degradasi Data Visual

Sumber: <https://www.mdpi.com/1424-8220/23/8/3784>

## 3.6 Mekanisme Monitoring

Mekanisme monitoring model dalam penelitian ini dirancang untuk mengamati perilaku model pembelajaran mesin secara berkelanjutan setelah model berada pada kondisi pasca-deployment. Monitoring difokuskan pada perubahan perilaku model akibat degradasi data operasional, tanpa melibatkan proses pelatihan ulang atau adaptasi model. Dengan pendekatan ini, mekanisme monitoring merepresentasikan praktik umum yang diterapkan pada sistem produksi yang bersifat stabil dan terkontrol.

Monitoring dilakukan secara batch-based, di mana data operasional dikelompokkan ke dalam batch yang merepresentasikan satu periode pengamatan tertentu. Setiap batch diperlakukan sebagai satu unit evaluasi independen, sehingga perubahan perilaku model dapat diamati secara temporal dan dibandingkan antar periode. Pendekatan ini dipilih untuk menyeimbangkan kebutuhan monitoring yang konsisten dengan keterbatasan evaluasi real-time di lingkungan produksi.



Gambar 3.5 Single vs Multi-Metrics Monitoring

Sumber: Penulis, 2025

3.6.1 Skema Monitoring Batch

Dalam skema monitoring batch, data operasional yang telah mengalami degradasi diproses secara bertahap sesuai dengan urutan batch yang telah ditentukan. Untuk setiap batch, model baseline melakukan inferensi terhadap seluruh data dalam batch tersebut, dan output yang dihasilkan dikumpulkan sebagai dasar perhitungan metrik monitoring.

Skema ini memungkinkan pengamatan perubahan metrik monitoring dari batch ke batch, sehingga pola degradasi model dapat dianalisis secara temporal. Selain itu, pendekatan batch-based memudahkan integrasi metrik statistik dan metrik berbasis perilaku prediksi model, karena perhitungan metrik dilakukan pada kumpulan data yang cukup representatif dalam setiap periode pengamatan.

Monitoring batch juga memberikan fleksibilitas dalam penyesuaian ukuran batch dan interval monitoring sesuai dengan kebutuhan operasional. Dalam penelitian ini, ukuran batch dan interval monitoring ditetapkan secara konsisten pada seluruh eksperimen untuk menjaga keterbandingan hasil antara pendekatan single-metric dan multi-metric.

3.6.2 Single-Metric Monitoring Berbasis Population Stability Index (PSI)

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.

3.6.3 Monitoring Berbasis Multi-Metrik (Composite Health Score)

Pendekatan monitoring berbasis multi-metrik dirancang untuk mengatasi keterbatasan pendekatan single-metric dengan mengintegrasikan beberapa indikator degradasi model yang merepresentasikan dimensi perubahan yang berbeda. Pada setiap batch data operasional, berbagai metrik monitoring dihitung secara paralel, mencakup metrik statistik distribusi data, metrik berbasis perilaku prediksi model, serta metrik performa sistem.

Nilai masing-masing metrik kemudian dinormalisasi agar berada pada skala yang sebanding, sehingga kontribusi setiap metrik terhadap evaluasi kondisi model dapat dikontrol secara adil. Setelah proses normalisasi, metrik-metrik tersebut digabungkan menggunakan kerangka Multi-Criteria Decision Model dengan metode agregasi Weighted Sum Model untuk menghasilkan suatu composite health score.

Composite health score digunakan sebagai indikator utama kondisi kesehatan model pada setiap batch data operasional. Dengan pendekatan ini, perubahan perilaku model tidak hanya dideteksi dari satu dimensi, tetapi dari kombinasi berbagai sinyal degradasi yang saling melengkapi. Pendekatan ini diharapkan dapat menghasilkan sinyal monitoring yang lebih stabil, informatif, dan relevan terhadap kebutuhan pengambilan keputusan operasional.

Monitoring berbasis multi-metrik selanjutnya dibandingkan dengan monitoring berbasis single-metric untuk mengevaluasi perbedaan kemampuan deteksi degradasi model, khususnya dalam hal sensitivitas terhadap degradasi gradual dan konsistensi sinyal antar skenario degradasi data.

Tabel 3.5 Definisi Metrik Monitoring

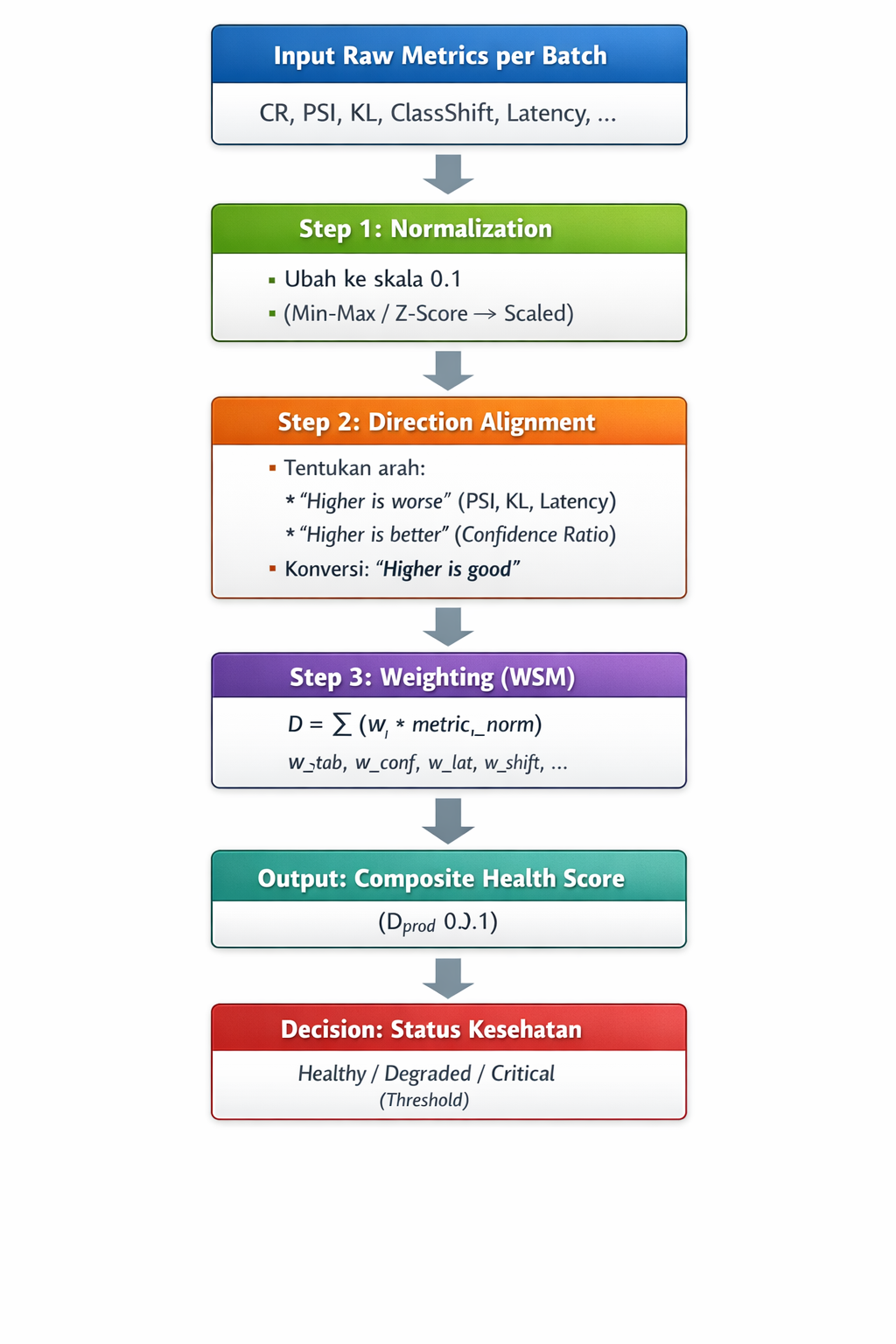
|  |  |  |
| --- | --- | --- |
| Metrik | Jenis | Fungsi |
| *Confidence Ratio* | *Single* | Kepercayaan prediksi |
| *PSI* | *Stability* | Pergeseran distribusi |
| *KL Divergence* | *Stability* | Divergensi probabilitas |
| *Class Shift* | *Distribution* | Perubahan proporsi kelas |
| *Latency* | Operational | Kinerja sistem |

## 3.7 Definisi, Normalisasi, dan Agregasi Metrik

Pendekatan monitoring berbasis multi-metrik dalam penelitian ini mengandalkan integrasi beberapa indikator degradasi model yang merepresentasikan dimensi perubahan yang berbeda. Agar integrasi tersebut dapat dilakukan secara adil dan terukur, setiap metrik perlu didefinisikan secara jelas, dinormalisasi ke dalam skala yang sebanding, dan digabungkan menggunakan metode agregasi yang terstruktur.

Dalam penelitian ini, composite health score tidak diposisikan sebagai sekadar hasil agregasi numerik dari beberapa metrik monitoring, melainkan sebagai decision-level indicator dalam proses monitoring model pasca-deployment. Skor komposit ini dirancang untuk merepresentasikan kondisi kesehatan model secara holistik dengan mengintegrasikan berbagai dimensi degradasi, meliputi perubahan distribusi data, stabilitas perilaku prediksi, serta performa sistem secara operasional. Dengan demikian, nilai composite health score digunakan sebagai dasar pengambilan keputusan operasional, seperti indikasi perlunya peningkatan pemantauan, intervensi sistem, atau rollback model, bukan hanya sebagai alat analisis pasca-kejadian

Subbab ini menjelaskan definisi metrik yang digunakan, skema normalisasi yang diterapkan, serta mekanisme agregasi metrik untuk membentuk composite health score sebagai indikator utama kondisi kesehatan model..



Gambar 3.6 Proses Normalisasi dan Agregasi Metrik menggunakan WSM

Sumber: Penulis, 2025

3.7.1 Definisi Metrik Monitoring

Metrik monitoring yang digunakan dalam penelitian ini dipilih untuk merepresentasikan berbagai dimensi degradasi model pembelajaran mesin di lingkungan produksi. Metrik-metrik tersebut mencakup indikator perubahan distribusi data, perubahan perilaku prediksi model, serta aspek performa sistem yang relevan secara operasional.

Metrik statistik digunakan untuk mendeteksi pergeseran distribusi data antara kondisi baseline dan kondisi operasional. Dalam penelitian ini, Population Stability Index (PSI) digunakan sebagai metrik utama untuk merepresentasikan perubahan distribusi data dan berfungsi sebagai baseline single-metric monitoring. Selain PSI, metrik statistik lain dapat digunakan untuk menangkap perubahan distribusi yang bersifat komplementer.

Metrik berbasis perilaku prediksi model digunakan untuk mengamati perubahan karakteristik output model, seperti pergeseran proporsi kelas prediksi atau perubahan tingkat kepercayaan prediksi. Metrik ini memberikan informasi tambahan mengenai bagaimana perubahan data memengaruhi perilaku internal model, yang tidak dapat ditangkap oleh metrik distribusi data semata.

Selain itu, metrik performa sistem digunakan untuk merepresentasikan dampak degradasi terhadap aspek operasional, seperti latency inferensi. Meskipun metrik ini tidak secara langsung merefleksikan kualitas prediksi model, perubahan signifikan pada performa sistem dapat menjadi indikator tambahan terhadap kondisi kesehatan model secara keseluruhan.

Pemilihan kombinasi metrik ini bertujuan untuk menghasilkan representasi kondisi kesehatan model yang lebih komprehensif dibandingkan penggunaan satu metrik tunggal.

3.7.2 Skema Normalisasi Metrik

Metrik-metrik monitoring yang digunakan dalam penelitian ini memiliki satuan dan rentang nilai yang berbeda-beda. Oleh karena itu, proses normalisasi diperlukan untuk memastikan bahwa setiap metrik berkontribusi secara proporsional dalam proses agregasi.

Normalisasi dilakukan dengan memetakan nilai masing-masing metrik ke dalam rentang skala yang sama, yaitu antara 0 dan 1. Dalam penelitian ini, metrik dibedakan menjadi dua kategori, yaitu metrik bertipe “baik” (good metric) dan metrik bertipe “buruk” (bad metric). Untuk metrik bertipe good, nilai yang lebih tinggi merepresentasikan kondisi model yang lebih baik, sedangkan untuk metrik bertipe bad, nilai yang lebih tinggi merepresentasikan kondisi model yang lebih buruk.

Rentang normalisasi untuk setiap metrik ditentukan berdasarkan nilai baseline dan hasil observasi awal eksperimen, bukan berdasarkan batas teoretis yang kaku. Pendekatan ini dipilih untuk memastikan bahwa normalisasi mencerminkan kondisi operasional yang realistis dan relevan dengan konteks eksperimen. Dengan demikian, perubahan kecil namun signifikan pada nilai metrik dapat terdeteksi secara lebih sensitif dalam composite health score. Hasil normalisasi menghasilkan nilai metrik terstandardisasi yang dapat dibandingkan dan digabungkan secara langsung pada tahap agregasi.

3.7.3 Perhitungan Composite Health Score

Setelah seluruh metrik dinormalisasi, tahap selanjutnya adalah agregasi metrik untuk membentuk satu indikator komposit yang merepresentasikan kondisi kesehatan model secara keseluruhan. Dalam penelitian ini, agregasi dilakukan menggunakan kerangka Multi-Criteria Decision Model dengan metode Weighted Sum Model (WSM).

Pada metode WSM, setiap metrik dinormalisasi dikalikan dengan bobot yang merepresentasikan tingkat kepentingan relatif metrik tersebut dalam konteks monitoring. Bobot ditentukan berdasarkan pertimbangan kebijakan monitoring dan relevansi operasional, dengan asumsi bahwa total bobot seluruh metrik bernilai satu. Hasil perkalian tersebut kemudian dijumlahkan untuk menghasilkan composite health score pada setiap batch data operasional.

Composite health score yang dihasilkan berada pada skala kontinu dan digunakan sebagai indikator utama untuk mengevaluasi kondisi kesehatan model dari waktu ke waktu. Nilai skor yang lebih tinggi mengindikasikan kondisi model yang lebih sehat, sedangkan penurunan skor mengindikasikan terjadinya degradasi model. Skor ini selanjutnya dianalisis untuk menentukan threshold operasional yang membedakan kondisi model yang sehat dan terdegradasi.

Dengan pendekatan ini, informasi multi-dimensi yang diperoleh dari berbagai metrik monitoring dapat disintesis menjadi satu sinyal yang mudah diinterpretasikan dan relevan untuk mendukung pengambilan keputusan operasional. Subbab berikutnya akan membahas proses penentuan threshold composite health score secara empiris melalui eksperimen terkontrol. Secara umum, *Composite Health Score D-Prod* dihitung menggunakan persamaan berikut:

Dₚᵣₒd = ∑ᵢ₌₁ⁿ wᵢ · mᵢ

dengan:

: bobot kriteria ke-i

: nilai metrik ke-i yang telah dinormalisasi

: 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.

Dalam penelitian ini, composite health score tidak diposisikan sebagai sekadar hasil agregasi numerik dari beberapa metrik monitoring, melainkan sebagai *decision-level indicator* dalam proses monitoring model pasca-deployment. Skor komposit ini dirancang untuk merepresentasikan kondisi kesehatan model secara holistik dengan mengintegrasikan berbagai dimensi degradasi, meliputi perubahan distribusi data, stabilitas perilaku prediksi, serta performa sistem secara operasional. Dengan demikian, nilai composite health score digunakan sebagai dasar pengambilan keputusan operasional, seperti indikasi perlunya peningkatan pemantauan, intervensi sistem, atau rollback model, bukan hanya sebagai alat analisis pasca-kejadian.

## 3.8 Penentuan Threshold Composite Health Score

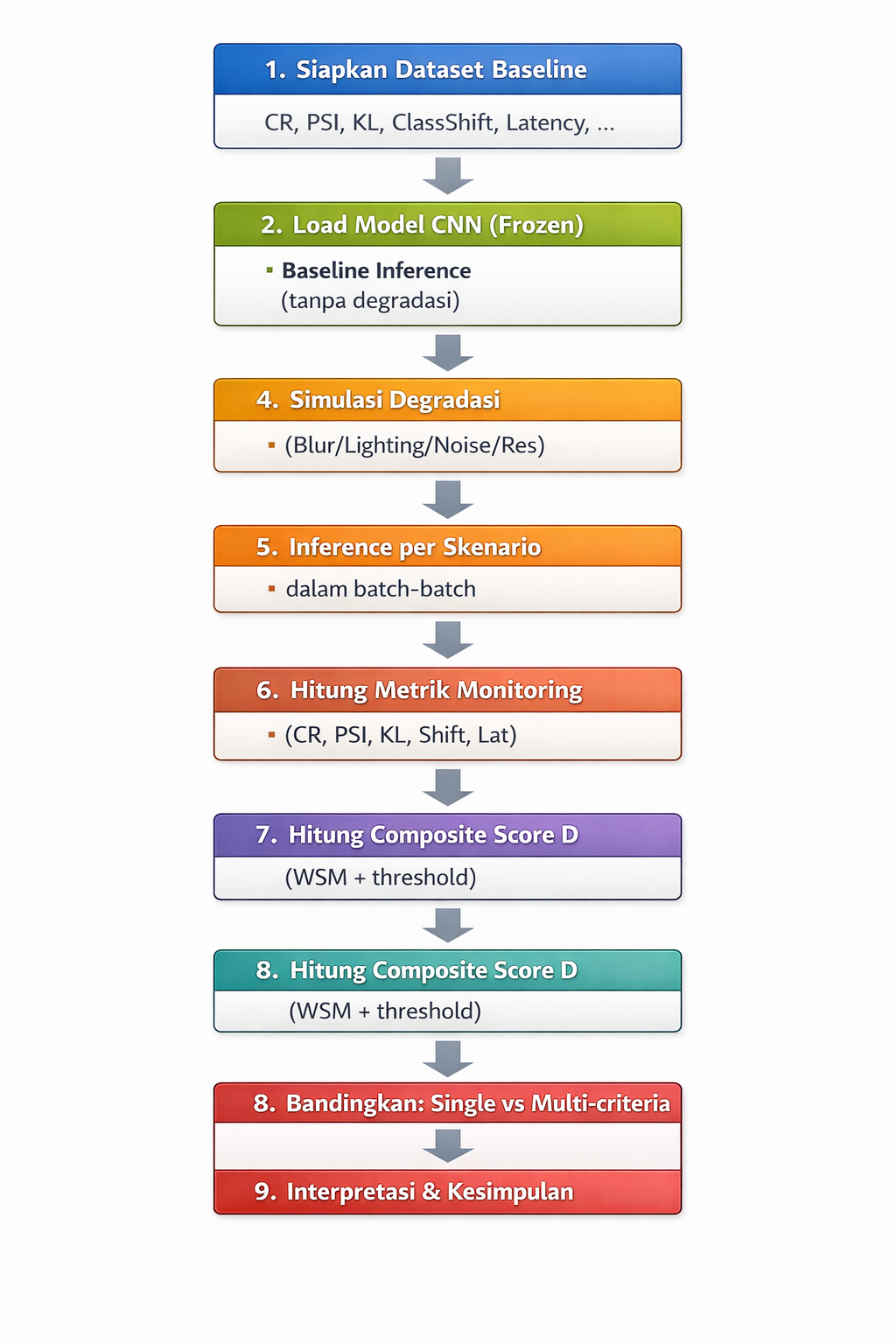
Penentuan threshold composite health score merupakan tahap penting dalam penerapan mekanisme monitoring model pembelajaran mesin di lingkungan produksi, karena threshold berfungsi sebagai batas keputusan yang membedakan kondisi model yang masih dianggap sehat dengan kondisi model yang terindikasi mengalami degradasi. Dalam konteks penelitian ini, threshold tidak ditetapkan secara heuristik, melainkan diturunkan secara empiris berdasarkan perilaku skor yang diamati selama eksperimen.

Penentuan bobot dan threshold pada composite health score dalam penelitian ini dilakukan dengan pendekatan *policy-based*, bukan melalui proses optimasi untuk mengejar hasil tertentu. Pendekatan ini dipilih untuk merefleksikan kondisi nyata di lingkungan produksi, di mana keputusan monitoring model umumnya ditentukan oleh kebijakan operasional, toleransi risiko, serta kebutuhan stabilitas sistem, bukan semata-mata oleh performa metrik individual. Oleh karena itu, threshold composite health score ditetapkan berdasarkan observasi empiris terhadap perilaku metrik pada skenario degradasi terkontrol, sehingga batasan yang dihasilkan bersifat interpretable, konsisten, dan aplikatif dalam konteks operasional.

Sebagaimana telah dijelaskan pada Subbab 3.5, mekanisme monitoring dalam penelitian ini dievaluasi pada lima kondisi skenario, yaitu satu kondisi baseline tanpa degradasi dan empat kondisi degradasi data (blur, compression, lighting, dan Gaussian noise). Evaluasi pada kelima skenario tersebut bertujuan untuk menilai sensitivitas dan stabilitas sinyal monitoring terhadap berbagai jenis gangguan kualitas data.

Namun, eksperimen penentuan threshold composite health score dibatasi pada dua skenario degradasi, yaitu blur dan compression. Pembatasan ini dilakukan untuk menjaga fokus analisis dan kedalaman eksperimen, serta karena kedua skenario tersebut merepresentasikan degradasi yang bersifat gradual dan umum dijumpai pada sistem computer vision di lingkungan produksi. Dengan demikian, threshold yang diturunkan diharapkan mencerminkan kondisi transisi yang relevan secara operasional, tanpa dipengaruhi oleh variasi degradasi yang bersifat ekstrem atau sangat kontekstual.

Pendekatan ini memungkinkan penentuan threshold dilakukan secara lebih terkontrol dan interpretatif, sekaligus menyediakan dasar yang jelas untuk mengevaluasi konsistensi dan robustnes composite health score sebagai indikator kondisi kesehatan model. Subbab berikutnya menguraikan rasional, desain eksperimen, serta hasil penentuan threshold pada masing-masing skenario degradasi yang digunakan.



Gambar 3.7 Prosedur Eksperimen

Sumber: Penulis, 2025

3.8.1 Rasional Penentuan Threshold

Pendekatan penentuan threshold dalam penelitian ini didasarkan pada asumsi bahwa degradasi model tidak terjadi secara tiba-tiba, melainkan berkembang secara bertahap seiring dengan penurunan kualitas data operasional. Oleh karena itu, threshold ditetapkan berdasarkan titik transisi perilaku composite health score yang menunjukkan perubahan signifikan dari kondisi stabil menuju kondisi terdegradasi.

Berbeda dengan pendekatan berbasis aturan statis atau nilai baku tertentu, penelitian ini menekankan penentuan threshold yang diturunkan dari observasi empiris terhadap sinyal monitoring. Dengan cara ini, threshold yang dihasilkan tidak hanya bersifat teoritis, tetapi juga mencerminkan dinamika sistem yang diamati selama eksperimen. Pendekatan ini juga memungkinkan threshold yang ditetapkan untuk bersifat lebih adaptif terhadap karakteristik metrik yang digunakan dan konteks operasional model, sehingga hasil monitoring menjadi lebih relevan untuk mendukung pengambilan keputusan.

3.8.2 Desain Eksperimen Penentuan Threshold

Eksperimen penentuan threshold dilakukan dengan menjalankan mekanisme monitoring berbasis multi-metrik pada data operasional yang mengalami degradasi secara bertahap. Monitoring dilakukan secara batch-based, di mana setiap batch merepresentasikan tingkat degradasi tertentu sesuai dengan skenario yang dirancang.

Pada setiap batch, composite health score dihitung dan dicatat untuk dianalisis secara temporal. Perubahan skor diamati untuk mengidentifikasi pola penurunan yang konsisten dan titik di mana skor mulai menyimpang secara signifikan dari kondisi baseline. Analisis ini dilakukan tanpa mengacu pada metrik performa berbasis ground truth, sehingga pendekatan tetap konsisten dengan kondisi produksi yang memiliki keterbatasan label.

Eksperimen penentuan threshold dibatasi pada dua skenario degradasi data, yaitu blur dan compression. Pembatasan ini dilakukan untuk menjaga fokus penelitian dan memastikan bahwa analisis threshold dilakukan secara mendalam dan terkontrol, tanpa menambah kompleksitas dari variasi skenario yang berlebihan.

3.8.3 Penentuan Threshold pada Skenario Degradasi Blur

Pada skenario degradasi blur, composite health score diamati pada setiap batch data dengan tingkat blur yang meningkat secara bertahap. Pada kondisi awal, skor menunjukkan nilai yang relatif stabil dan merepresentasikan kondisi model yang sehat. Seiring dengan peningkatan intensitas blur, skor mulai menunjukkan tren penurunan yang konsisten.

Threshold ditentukan pada titik di mana penurunan skor mulai menunjukkan perubahan pola yang signifikan dan berkelanjutan, yang mengindikasikan bahwa degradasi data telah mulai berdampak secara nyata terhadap perilaku model. Titik ini dipilih sebagai batas antara kondisi model yang masih dapat ditoleransi dan kondisi yang memerlukan perhatian atau intervensi lebih lanjut. Pendekatan ini memungkinkan threshold untuk merepresentasikan kondisi transisi yang relevan secara operasional, bukan sekadar perubahan statistik yang tidak berdampak langsung terhadap kesehatan model.

3.8.4 Penentuan Threshold pada Skenario Degradasi Compression

Pada skenario degradasi compression, eksperimen dilakukan dengan pendekatan yang sama seperti pada skenario blur. Composite health score dihitung pada setiap batch data dengan tingkat kompresi yang meningkat secara bertahap. Pola perubahan skor dianalisis untuk mengidentifikasi konsistensi perilaku skor terhadap degradasi jenis yang berbeda.

Hasil observasi pada skenario compression digunakan untuk mengevaluasi apakah threshold yang ditentukan pada skenario blur menunjukkan konsistensi atau memerlukan penyesuaian. Konsistensi threshold antar skenario degradasi menjadi indikator penting terhadap robustnes pendekatan monitoring berbasis multi-metrik yang diusulkan.

Dengan menganalisis dua skenario degradasi yang berbeda, penelitian ini mengevaluasi kemampuan composite health score dalam memberikan sinyal monitoring yang stabil dan dapat diandalkan, serta memastikan bahwa threshold yang ditetapkan tidak bersifat terlalu spesifik terhadap satu jenis degradasi tertentu.

## 3.9 Kriteria Evaluasi dan Analisis Perbandingan

Prosedur evaluasi dalam penelitian ini dirancang untuk membandingkan efektivitas pendekatan monitoring berbasis single-metric dan multi-metrik dalam mendeteksi degradasi model pembelajaran mesin di lingkungan produksi. Evaluasi difokuskan pada kemampuan masing-masing pendekatan dalam memberikan sinyal monitoring yang relevan, stabil, dan dapat digunakan sebagai dasar pengambilan keputusan operasional.

Perbandingan dilakukan dengan menerapkan kedua pendekatan monitoring pada skema eksperimen dan data operasional yang sama. Dengan pendekatan ini, perbedaan hasil evaluasi yang diperoleh dapat dikaitkan secara langsung dengan mekanisme monitoring yang digunakan, bukan dengan perbedaan kondisi eksperimen atau skema pengamatan.

Salah satu aspek utama yang dievaluasi adalah waktu deteksi degradasi. Waktu deteksi didefinisikan sebagai periode atau batch pertama di mana sinyal monitoring menunjukkan indikasi degradasi yang melewati threshold yang telah ditetapkan. Pendekatan yang mampu mendeteksi degradasi lebih awal dianggap lebih efektif dalam mendukung tindakan preventif di lingkungan produksi.

Aspek evaluasi berikutnya adalah stabilitas sinyal monitoring. Stabilitas diukur dengan mengamati fluktuasi nilai metrik atau skor monitoring antar batch data operasional. Pendekatan monitoring yang menghasilkan sinyal yang lebih konsisten dan tidak mudah terpengaruh oleh variasi kecil pada data dianggap lebih andal untuk digunakan secara operasional.

Selain itu, evaluasi juga mempertimbangkan relevansi operasional dari sinyal monitoring yang dihasilkan. Sinyal dianggap relevan apabila perubahan nilai metrik atau skor monitoring dapat diinterpretasikan secara jelas sebagai indikasi kondisi model yang sehat atau terdegradasi, serta dapat digunakan untuk mendukung keputusan seperti peningkatan pemantauan atau intervensi sistem.

Hasil monitoring berbasis single-metric dan multi-metrik dianalisis secara komparatif untuk mengidentifikasi perbedaan pola deteksi degradasi pada masing-masing skenario degradasi data. Analisis ini dilakukan secara deskriptif dan kuantitatif, dengan menitikberatkan pada konsistensi hasil antar skenario dan kesesuaian sinyal monitoring dengan kondisi degradasi yang dirancang.

Melalui prosedur evaluasi dan perbandingan ini, penelitian ini bertujuan untuk menunjukkan secara empiris keunggulan dan keterbatasan masing-masing pendekatan monitoring. Hasil evaluasi selanjutnya disajikan dan dibahas secara rinci pada Bab IV untuk mendukung kesimpulan penelitian.

## 3.10 Validitas dan Keterbatasan Metodologi

Setiap penelitian eksperimental memiliki keterbatasan yang perlu diakui untuk menjaga objektivitas dan validitas hasil yang diperoleh. Pada subbab ini dibahas aspek validitas metodologi penelitian serta keterbatasan yang melekat pada desain eksperimen yang digunakan.

Dari sisi validitas internal, penelitian ini dirancang dengan skema eksperimen yang terkontrol dan konsisten. Model baseline yang digunakan bersifat tetap selama seluruh eksperimen, sehingga perubahan perilaku model yang teramati dapat dikaitkan secara langsung dengan degradasi data operasional yang disimulasikan. Selain itu, penggunaan skema monitoring batch yang sama pada pendekatan single-metric dan multi-metrik memastikan bahwa perbandingan hasil monitoring dilakukan secara adil dan objektif.

Namun demikian, terdapat keterbatasan pada validitas eksternal penelitian ini. Eksperimen dilakukan pada satu model pembelajaran mesin dan satu jenis tugas klasifikasi berbasis computer vision, sehingga generalisasi hasil penelitian ke domain atau arsitektur model lain perlu dilakukan dengan kehati-hatian. Hasil yang diperoleh mencerminkan perilaku monitoring pada konteks dan konfigurasi eksperimen yang digunakan, dan tidak secara langsung menjamin kinerja yang sama pada sistem dengan karakteristik yang berbeda.

Keterbatasan lainnya berkaitan dengan skenario degradasi data yang digunakan. Penelitian ini memfokuskan eksperimen penentuan threshold pada dua jenis degradasi, yaitu blur dan compression. Pembatasan ini dilakukan untuk menjaga fokus dan kedalaman analisis, namun implikasinya adalah threshold yang diperoleh belum diuji secara menyeluruh terhadap jenis degradasi lain, seperti variasi pencahayaan atau noise. Oleh karena itu, penerapan threshold pada konteks yang lebih luas memerlukan evaluasi tambahan.

Dari sisi pemilihan metrik dan bobot, penentuan bobot pada pendekatan multi-metrik didasarkan pada pertimbangan kebijakan monitoring dan relevansi operasional. Meskipun pendekatan ini mencerminkan praktik yang realistis di lingkungan produksi, pemilihan bobot tetap mengandung unsur subjektivitas. Penelitian ini belum mengeksplorasi secara mendalam sensitivitas hasil terhadap variasi bobot metrik, sehingga aspek tersebut dapat menjadi arah penelitian lanjutan.

Selain itu, penelitian ini tidak mencakup integrasi mekanisme monitoring ke dalam pipeline MLOps secara end-to-end, seperti otomatisasi tindakan rollback, alerting, atau retraining. Fokus penelitian dibatasi pada evaluasi metodologi monitoring dan analisis eksperimen, sehingga implementasi operasional penuh pada sistem produksi nyata berada di luar ruang lingkup penelitian ini.

Dengan mempertimbangkan validitas dan keterbatasan tersebut, hasil penelitian ini diharapkan dapat dipahami secara proporsional dan digunakan sebagai dasar untuk pengembangan lebih lanjut. Pendekatan metodologis yang diusulkan memberikan kontribusi pada pemahaman mengenai monitoring degradasi model pembelajaran mesin berbasis multi-metrik, sekaligus membuka peluang penelitian lanjutan untuk memperluas cakupan dan penerapan metode yang dikembangkan.

# BAB IV

# HASIL PENELITIAN DAN PEMBAHASAN

## 4.1 Gambaran Umum Pelaksanaan Eksperimen

Bab ini menyajikan hasil penelitian dan pembahasan berdasarkan eksperimen yang telah dirancang pada Bab III. Fokus utama Bab IV adalah mengevaluasi perilaku model pembelajaran mesin pada fase pasca-deployment melalui pendekatan monitoring model, tanpa melibatkan proses pelatihan ulang (retraining) maupun perubahan arsitektur model.

Eksperimen dilakukan menggunakan model klasifikasi citra berbasis computer vision yang telah dilatih sebelumnya dan diperlakukan sebagai frozen model. Evaluasi dilakukan pada lingkungan produksi tersimulasikan (production-like environment) dengan skema monitoring berbasis batch (batch-based monitoring), yang merepresentasikan praktik umum pemantauan model dalam sistem MLOps.

Data yang digunakan dalam eksperimen terdiri atas kondisi baseline tanpa degradasi serta beberapa skenario degradasi data visual yang dirancang secara sintetis. Skenario degradasi tersebut meliputi blur, compression (JPEG), lighting variation, dan Gaussian noise. Seluruh skenario diterapkan menggunakan parameter transformasi yang telah ditetapkan pada tahap metodologi penelitian dan digunakan secara konsisten pada seluruh eksperimen.

Dalam pelaksanaannya, dua pendekatan monitoring dibandingkan, yaitu pendekatan single-metric monitoring dan pendekatan multi-metric monitoring. Pendekatan single-metric menggunakan satu indikator utama untuk mendeteksi degradasi model, sedangkan pendekatan multi-metric mengamati kondisi model dari berbagai dimensi melalui kombinasi beberapa metrik monitoring. Selanjutnya, hasil dari pendekatan multi-metric diagregasikan ke dalam satu nilai komposit yang merepresentasikan tingkat kesehatan model.

Evaluasi pada Bab IV difokuskan pada analisis perubahan sinyal monitoring, stabilitas indikator, serta kemampuan masing-masing pendekatan dalam mendeteksi degradasi model pada berbagai skenario data. Dengan demikian, pembahasan pada bab ini menitikberatkan pada perilaku dan interpretasi hasil monitoring, bukan pada perbandingan performa klasifikasi secara konvensional.

Tabel 4.1 – Ringkasan Setup Eksperimen

|  |  |
| --- | --- |
| **Komponen** | **Deskripsi** |
| **Jenis Penelitian** | Eksperimen komparatif (*comparative experimental study*) |
| **Domain Aplikasi** | *Computer vision* – klasifikasi citra |
| **Tugas Klasifikasi** | Klasifikasi biner (dua kelas) |
| **Dataset** | Dataset citra *computer vision* dengan skenario baseline dan degradasi sintetis |
| **Model Baseline** | Convolutional Neural Network (CNN) – MobileNetV3 |
| **Status Model** | *Frozen model* (tanpa retraining selama eksperimen) |
| **Lingkungan Evaluasi** | Lingkungan produksi tersimulasikan (*production-like environment*) |
| **Mode Monitoring** | *Batch-based monitoring* |
| **Pendekatan Monitoring** | Single-metric monitoring dan multi-metric monitoring |
| **Single-Metric** | Population Stability Index (PSI) |
| **Multi-Metrics** | PSI, KL Divergence, Class Shift, Confidence Ratio, Latency (P95) |
| **Metode Agregasi** | Multi-Criteria Decision Model (MCDM) dengan Weighted Sum Model (WSM) |
| **Normalisasi Metrik** | Normalisasi eksplisit berbasis kategori *good metrics* dan *bad metrics* |
| **Output Monitoring** | Composite health score (nilai kesehatan model) |
| **Skenario Data** | Baseline (tanpa degradasi) dan data terdegradasi |
| **Jenis Degradasi Data** | Blur, Compression (JPEG) |
| **Unit Analisis** | Batch data operasional per skenario degradasi |
| **Tujuan Evaluasi** | Membandingkan efektivitas single-metric dan multi-metric monitoring dalam mendeteksi degradasi model |
| **Hasil yang Dianalisis** | Sensitivitas deteksi degradasi, stabilitas sinyal monitoring, dan interpretabilitas kondisi model |

## 4.2 Hasil Baseline Model pada Kondisi Normal

Subbab ini menyajikan hasil evaluasi model pada kondisi baseline, yaitu kondisi data operasional tanpa penerapan skenario degradasi. Kondisi baseline digunakan sebagai titik acuan (reference state) untuk menilai perubahan perilaku model pada skenario degradasi yang dianalisis pada subbab selanjutnya.

Pada kondisi baseline, model menunjukkan perilaku inferensi yang stabil baik dari sisi distribusi output prediksi maupun metrik monitoring yang diamati. Nilai metrik monitoring yang dihitung pada fase ini mencerminkan kesesuaian antara data operasional dan data referensi, sehingga tidak mengindikasikan adanya pergeseran distribusi data atau anomali perilaku model.

Selain itu, indikator perilaku model seperti confidence ratio dan latency inferensi berada pada rentang yang konsisten antar batch data. Kondisi ini mengindikasikan bahwa model beroperasi dalam keadaan sehat (healthy state) ketika tidak dihadapkan pada penurunan kualitas data input.

Hasil baseline ini selanjutnya digunakan sebagai dasar dalam proses normalisasi metrik monitoring serta sebagai pembanding utama untuk menginterpretasikan perubahan sinyal monitoring pada setiap skenario degradasi. Dengan menetapkan kondisi baseline yang stabil, analisis degradasi model pada subbab berikutnya dapat difokuskan secara spesifik pada dampak skenario degradasi data, bukan pada fluktuasi alami sistem.

Tabel 4.2 Performa Baseline Model (Tanpa Degradasi)

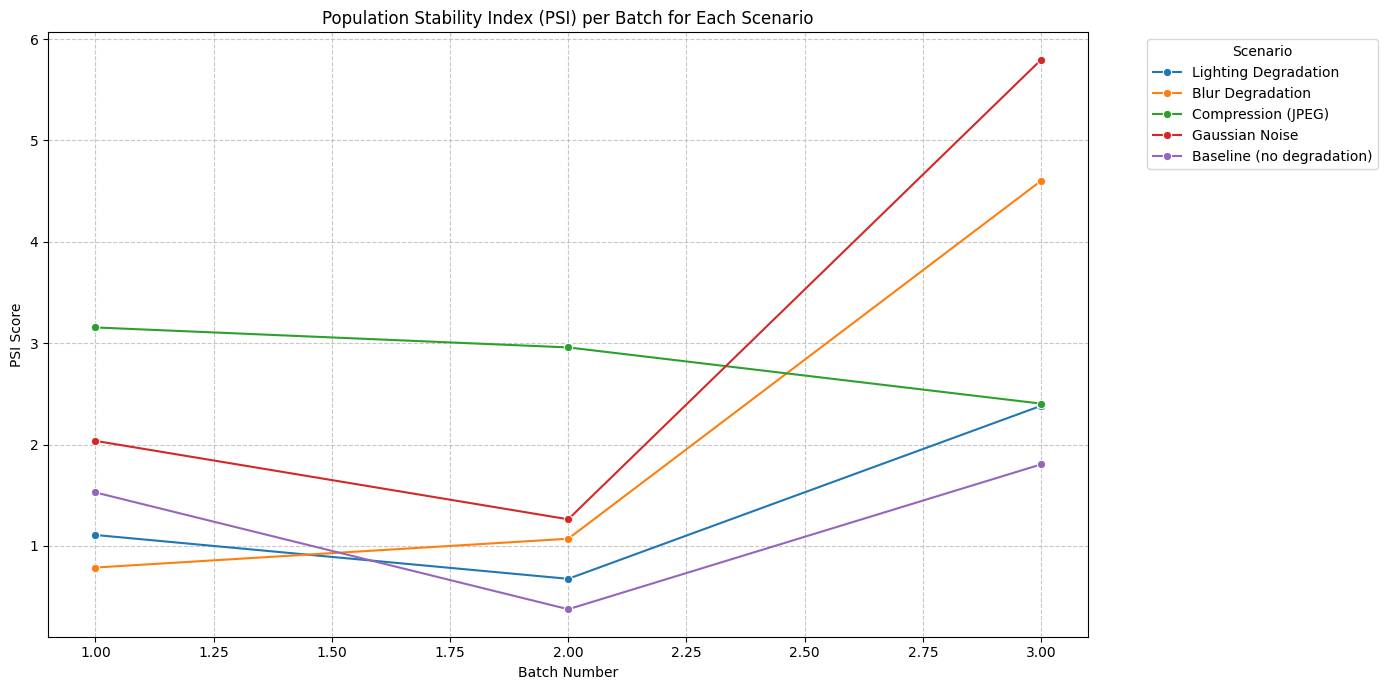
|  |  |
| --- | --- |
| **Metrik** | **Nilai** |
| Accuracy | 0.8158 |
| Prec | 0.8330 |
| Recall | 0.8158 |
| f1 | 0.7926 |
| PSI | 0.000000 |
| KL Divergence | 0.000000 |
| Class Shift | 0.000000 |
| Confidence Ratio | 1.0 |
| P95 Latency | 1342.019820 |

## 4.3 Hasil Monitoring pada Berbagai Skenario Degradasi

Subbab ini membahas hasil perbandingan antara pendekatan monitoring berbasis single-metric dan multi-metric dalam mendeteksi degradasi model pembelajaran mesin di lingkungan produksi. Analisis dilakukan dengan mensintesis hasil monitoring pada seluruh skenario degradasi yang telah dibahas pada Subbab 4.2 dan 4.3, dengan fokus pada karakteristik sinyal monitoring yang dihasilkan dan relevansinya terhadap pengambilan keputusan operasional.

Pendekatan single-metric dalam penelitian ini menggunakan Population Stability Index (PSI) sebagai indikator utama perubahan distribusi data. Hasil eksperimen menunjukkan bahwa PSI mampu mendeteksi adanya pergeseran distribusi data pada beberapa skenario degradasi, khususnya pada degradasi yang secara langsung memengaruhi distribusi nilai piksel. Namun, pada praktiknya sinyal PSI menunjukkan keterbatasan dalam memberikan gambaran kondisi kesehatan model secara konsisten.

Pada beberapa skenario degradasi, seperti lighting dan Gaussian noise, nilai PSI menunjukkan fluktuasi yang cukup tinggi antar batch, meskipun perubahan kualitas data berlangsung secara gradual. Fluktuasi ini berpotensi menghasilkan sinyal false alarm atau kesulitan dalam menentukan kapan degradasi telah mencapai tingkat yang berdampak signifikan terhadap perilaku model. Selain itu, sebagai metrik statistik tunggal, PSI tidak memberikan informasi mengenai bagaimana perubahan distribusi data tersebut memengaruhi perilaku prediksi model atau aspek operasional lainnya.



Gambar 4.1 – PSI vs Batch / Waktu

Sumber: Pribadi, 2026

Tabel 4.3 – Nilai PSI per Skenario Degradasi

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Scenario | **Mean Max Conf** | **Overall PSI** | **P95 Latency (ms)** | **Num Batches** | **Num Samples** |
| Baseline (no degradation) | 0.958844 | 0.000000 | 2301.311469 | 3 | 76 |
| Lighting Degradation | 0.958801 | 0.385190 | 2682.917356 | 3 | 76 |
| Blur Degradation | 0.922971 | 0.952615 | 1347.087359 | 3 | 76 |
| Gaussian Noise | 0.908480 | 1.041468 | 2544.256020 | 3 | 76 |
| Compression (JPEG) | 0.924384 | 1.550903 | 1816.796947 | 3 | 76 |

Sebaliknya, pendekatan monitoring berbasis multi-metrik menunjukkan karakteristik sinyal yang lebih stabil dan informatif pada seluruh skenario degradasi. Composite health score yang dihasilkan mampu merepresentasikan akumulasi dampak degradasi data terhadap berbagai dimensi monitoring secara simultan. Integrasi beberapa metrik memungkinkan perubahan kecil pada satu dimensi tidak secara langsung mendominasi sinyal monitoring secara keseluruhan, sehingga fluktuasi ekstrem dapat diredam.

Dari sisi waktu deteksi degradasi, pendekatan multi-metrik menunjukkan kecenderungan untuk memberikan indikasi penurunan kondisi model secara lebih gradual dan konsisten. Meskipun pendekatan single-metric dapat mendeteksi pergeseran distribusi pada tahap awal tertentu, sinyal yang dihasilkan sering kali sulit diinterpretasikan secara operasional karena tidak mencerminkan dampak degradasi terhadap perilaku model secara menyeluruh. Composite health score memberikan sinyal yang lebih mudah ditelusuri secara temporal dan lebih relevan untuk digunakan sebagai dasar pengambilan keputusan.

Dari perspektif stabilitas sinyal monitoring, pendekatan multi-metrik juga menunjukkan keunggulan. Pada skenario degradasi yang bersifat stokastik, seperti Gaussian noise, sinyal PSI cenderung menunjukkan variabilitas yang tinggi antar batch. Sebaliknya, composite health score tetap menunjukkan pola perubahan yang lebih terkontrol, sehingga lebih sesuai untuk diterapkan pada sistem monitoring produksi yang membutuhkan sinyal yang andal dan tidak mudah terpengaruh oleh variasi acak.

Analisis ini menunjukkan bahwa penggunaan satu metrik statistik tunggal, meskipun sederhana dan mudah diimplementasikan, memiliki keterbatasan dalam merepresentasikan kondisi kesehatan model secara komprehensif. Pendekatan multi-metrik menawarkan keunggulan dengan menggabungkan berbagai sinyal degradasi ke dalam satu indikator yang lebih stabil dan interpretatif. Dengan demikian, pendekatan monitoring berbasis multi-metrik lebih sesuai untuk mendukung kebutuhan monitoring model pembelajaran mesin di lingkungan produksi yang kompleks dan dinamis.

Hasil perbandingan ini menjadi dasar bagi pembahasan selanjutnya terkait penentuan dan validasi threshold composite health score, yang dibahas secara khusus pada Subbab 4.5

Tabel 4.3 Ringkasan Nilai Multi-Metrics per Skenario

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Scenario** | **PSI** | **KL** | **Class Shift** | **Conf Ratio** | **P95 Lat (ms)** | **Optimized D\_prod** |
| Baseline | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 1342.019820 | 0.991267 |
| Compression (JPEG) | 0.753538 | 0.054866 | 0.078947 | 0.973832 | 700.204945 | 0.849667 |
| Blur Degradation | 1.125034 | 0.734624 | 0.144737 | 1.022961 | 732.148743 | 0.650701 |
| Lighting Degradation | 1.762810 | 1.688077 | 0.144737 | 0.954980 | 880.019951 | 0.433725 |
| Gaussian Noise | 2.715236 | 0.480327 | 0.315789 | 0.857140 | 2317.536736 | 0.363304 |

4.3.1 Skenario Degradasi Blur

Pada skenario blur, degradasi data diterapkan untuk mensimulasikan kondisi penurunan ketajaman citra yang umum terjadi pada sistem computer vision di lingkungan produksi, seperti akibat gerakan kamera, fokus yang tidak optimal, atau kualitas sensor yang menurun. Penerapan degradasi blur bertujuan untuk mengevaluasi sensitivitas model terhadap perubahan kualitas visual yang berdampak langsung pada ekstraksi fitur.

Hasil monitoring menunjukkan bahwa penerapan blur memengaruhi perilaku model secara bertahap seiring meningkatnya tingkat degradasi. Perubahan sinyal monitoring mulai terlihat pada beberapa batch data, mengindikasikan adanya pergeseran karakteristik input yang diproses oleh model. Namun, pada tingkat blur tertentu, perubahan tersebut belum selalu diikuti oleh penurunan performa prediksi yang signifikan, sehingga interpretasi kondisi model memerlukan pengamatan yang lebih kontekstual.

Selain itu, sinyal monitoring pada skenario blur menunjukkan pola perubahan yang relatif halus dan progresif dibandingkan dengan skenario degradasi lainnya. Karakteristik ini mencerminkan sifat degradasi blur yang cenderung bersifat gradual, sehingga deteksi degradasi model pada skenario ini menuntut indikator monitoring yang mampu menangkap perubahan kecil namun konsisten.

Temuan pada skenario blur menjadi dasar penting dalam mengevaluasi kemampuan pendekatan monitoring untuk mendeteksi degradasi model secara dini. Analisis lebih lanjut mengenai respons model terhadap skenario blur, termasuk perbandingan antara pendekatan single-metric dan multi-metric, akan dibahas pada subbab selanjutnya.

4.3.2 Skenario Degradasi Compression

Pada skenario compression (JPEG), degradasi data diterapkan untuk mensimulasikan penurunan kualitas citra akibat proses kompresi lossy yang umum digunakan dalam sistem penyimpanan dan transmisi data visual. Kompresi citra sering dilakukan untuk mengurangi ukuran data, namun dapat menghilangkan detail visual penting yang berpengaruh terhadap representasi fitur yang dipelajari oleh model.

Hasil monitoring menunjukkan bahwa penerapan kompresi JPEG memengaruhi perilaku model secara lebih diskrit dibandingkan dengan skenario blur. Pada tingkat kompresi tertentu, perubahan sinyal monitoring muncul secara lebih jelas pada beberapa batch data, mencerminkan adanya perubahan karakteristik visual yang signifikan akibat hilangnya informasi frekuensi tinggi pada citra.

Selain itu, respons model pada skenario compression cenderung menunjukkan ambang perubahan yang lebih tegas, di mana kondisi model dapat beralih dari keadaan stabil ke kondisi terdegradasi dalam rentang batch yang relatif lebih sempit. Karakteristik ini menunjukkan bahwa degradasi akibat kompresi dapat berdampak langsung terhadap kemampuan model dalam mempertahankan konsistensi prediksi.

Temuan pada skenario compression menegaskan pentingnya mekanisme monitoring yang mampu menangkap perubahan kualitas data yang bersifat non-gradual. Analisis lebih lanjut mengenai efektivitas pendekatan monitoring dalam mendeteksi degradasi akibat kompresi, termasuk perbandingan antara pendekatan single-metric dan multi-metric, akan dibahas pada subbab berikutnya.

4.3.3 Skenario Degradasi Lighting

Pada skenario *lighting variation*, degradasi data diterapkan untuk mensimulasikan perubahan kondisi pencahayaan yang umum terjadi pada lingkungan operasional sistem *computer vision*. Variasi pencahayaan dapat disebabkan oleh perubahan intensitas cahaya, bayangan, atau perbedaan kondisi lingkungan, yang secara langsung memengaruhi distribusi nilai piksel pada citra input.

Hasil monitoring menunjukkan bahwa variasi pencahayaan memengaruhi perilaku model secara global, terutama pada perubahan distribusi intensitas citra. Perubahan tersebut dapat muncul meskipun struktur visual objek dalam citra relatif tidak berubah. Kondisi ini menyebabkan pergeseran karakteristik input yang diproses oleh model, yang tercermin pada sinyal monitoring pada beberapa batch data.

Berbeda dengan skenario blur dan compression yang memengaruhi detail visual atau informasi frekuensi tinggi, degradasi pencahayaan cenderung berdampak pada keseluruhan citra. Dampak ini dapat menghasilkan perubahan sinyal monitoring yang terlihat signifikan meskipun kualitas visual secara subjektif masih dapat diterima. Hal ini menunjukkan bahwa perubahan pencahayaan berpotensi memicu indikasi degradasi pada metrik tertentu tanpa selalu diikuti oleh penurunan performa prediksi yang langsung terlihat.

Temuan pada skenario lighting variation menekankan perlunya interpretasi sinyal monitoring yang mempertimbangkan konteks degradasi data. Analisis lanjutan terhadap respons model pada skenario ini, khususnya dalam membedakan perubahan distribusi data dan degradasi perilaku model yang sesungguhnya, akan dibahas pada subbab berikutnya

.Tabel 4.2 Status Model Berdasarkan *Single*-*Metric* Monitoring

|  |  |
| --- | --- |
| Skenario | Batch Deteksi |
| Lighting Degradation | 0.385190 |
| Blur Degradation | 0.952615 |
| Compression (JPEG) | 1.550903 |
| Noise | |  | | --- | | 1.041468 | |

4.3.4 Skenario Degradasi Gaussian Noise

Pada skenario Gaussian noise, degradasi data diterapkan untuk mensimulasikan gangguan acak pada citra input yang umum terjadi akibat keterbatasan sensor, interferensi lingkungan, atau proses akuisisi data yang tidak stabil. Penambahan noise acak menyebabkan fluktuasi nilai piksel yang bersifat tidak terstruktur, sehingga berpotensi mengganggu representasi fitur yang dipelajari oleh model.

Hasil monitoring menunjukkan bahwa penerapan Gaussian noise menghasilkan pola perubahan sinyal yang cenderung tidak konsisten antar batch data. Berbeda dengan skenario degradasi lain yang memiliki karakteristik perubahan yang lebih terarah, noise acak dapat memunculkan fluktuasi sinyal monitoring yang bersifat sporadis. Kondisi ini menyulitkan interpretasi degradasi model apabila hanya mengandalkan satu indikator monitoring.

Selain itu, dampak Gaussian noise terhadap perilaku model dapat bervariasi tergantung pada tingkat noise yang diterapkan. Pada beberapa batch, sinyal monitoring menunjukkan indikasi perubahan yang signifikan, sementara pada batch lainnya perubahan tersebut relatif kecil. Variabilitas ini mencerminkan sifat degradasi noise yang tidak deterministik dan menuntut mekanisme monitoring yang mampu membedakan fluktuasi acak dari degradasi model yang berkelanjutan.

Temuan pada skenario Gaussian noise melengkapi pemahaman mengenai respons model terhadap berbagai jenis degradasi data visual. Hasil ini menjadi dasar untuk pembahasan lebih lanjut mengenai efektivitas pendekatan monitoring dalam mengelola sinyal yang fluktuatif dan kompleks, yang akan dianalisis pada subbab berikutnya.

## 4.4 Analisis Monitoring Berdasarkan Pendekatan Single-Metric

Subbab ini membahas hasil monitoring model menggunakan pendekatan single-metric, di mana kondisi kesehatan model dievaluasi berdasarkan satu indikator utama. Pendekatan ini digunakan sebagai baseline pembanding untuk menilai efektivitas metode monitoring yang lebih kompleks pada subbab selanjutnya.

Pada pendekatan single-metric, sinyal monitoring dianalisis dengan mengamati perubahan nilai metrik terhadap kondisi baseline. Hasil eksperimen menunjukkan bahwa pendekatan ini mampu mendeteksi adanya perubahan distribusi data pada beberapa skenario degradasi. Namun, interpretasi kondisi model sering kali bergantung pada sensitivitas metrik yang digunakan serta karakteristik degradasi yang terjadi.

Pada skenario degradasi yang bersifat gradual, seperti blur, perubahan sinyal single-metric cenderung muncul secara perlahan. Kondisi ini berpotensi menyebabkan keterlambatan dalam mendeteksi degradasi model, terutama ketika perubahan kualitas data belum cukup signifikan untuk melampaui ambang batas yang ditetapkan. Sebaliknya, pada skenario degradasi yang bersifat lebih diskrit, seperti compression (JPEG), perubahan sinyal single-metric dapat muncul secara lebih jelas, meskipun tidak selalu mencerminkan dampak langsung terhadap perilaku prediksi model.

Selain itu, hasil monitoring menunjukkan bahwa pendekatan single-metric rentan menghasilkan sinyal yang ambigu pada skenario degradasi tertentu. Perubahan nilai metrik dapat mengindikasikan adanya pergeseran distribusi data, namun tidak selalu berkorelasi dengan penurunan kualitas prediksi atau stabilitas model secara keseluruhan. Kondisi ini menyulitkan pengambilan keputusan operasional apabila satu metrik digunakan sebagai satu-satunya indikator kesehatan model.

Berdasarkan hasil tersebut, pendekatan single-metric memiliki keterbatasan dalam merepresentasikan kondisi model secara komprehensif, khususnya pada lingkungan produksi yang dinamis dan kompleks. Temuan ini menjadi dasar untuk mengevaluasi pendekatan monitoring alternatif yang mempertimbangkan berbagai dimensi degradasi model secara simultan, sebagaimana dibahas pada subbab berikutnya.

4.4.1 Respons Single-Metric terhadap Skenario Degradasi

Subbab ini membahas respons monitoring model berdasarkan pendekatan *single-metric* terhadap berbagai skenario degradasi data yang telah dianalisis pada subbab sebelumnya. Evaluasi difokuskan pada pola perubahan sinyal monitoring yang dihasilkan ketika model dihadapkan pada penurunan kualitas data input.

Hasil monitoring menunjukkan bahwa pendekatan single-metric mampu menangkap adanya perubahan distribusi data pada beberapa skenario degradasi. Pada skenario degradasi yang bersifat gradual, seperti *blur*, perubahan sinyal cenderung muncul secara perlahan dan bertahap. Kondisi ini mencerminkan sensitivitas single-metric terhadap pergeseran distribusi data yang terjadi secara progresif.

Sebaliknya, pada skenario degradasi yang bersifat lebih diskrit, seperti *compression (JPEG)*, perubahan sinyal single-metric dapat muncul secara lebih jelas dalam rentang batch yang relatif singkat. Meskipun demikian, perubahan nilai metrik tidak selalu berkorelasi langsung dengan perubahan perilaku prediksi model, sehingga interpretasi kondisi kesehatan model masih memerlukan kehati-hatian.

Respons single-metric pada skenario degradasi lain, seperti *lighting variation* dan *Gaussian noise*, menunjukkan pola yang beragam. Pada kondisi tertentu, fluktuasi sinyal dapat muncul tanpa diikuti oleh indikasi degradasi model yang konsisten. Hal ini menandakan bahwa respons single-metric sangat dipengaruhi oleh karakteristik degradasi data yang terjadi.

4.4.2 Keterbatasan Pendekatan Single-Metric

Berdasarkan hasil eksperimen, pendekatan single-metric memiliki beberapa keterbatasan dalam merepresentasikan kondisi kesehatan model secara menyeluruh. Ketergantungan pada satu indikator utama menyebabkan interpretasi kondisi model sangat dipengaruhi oleh sensitivitas dan karakteristik metrik yang digunakan.

Salah satu keterbatasan utama adalah potensi ambiguitas sinyal monitoring. Perubahan nilai single-metric dapat mengindikasikan adanya pergeseran distribusi data, namun tidak selalu mencerminkan degradasi perilaku prediksi model yang signifikan. Kondisi ini menyulitkan pengambilan keputusan operasional apabila satu metrik digunakan sebagai satu-satunya dasar penilaian kesehatan model.

Selain itu, pendekatan single-metric cenderung kurang adaptif terhadap skenario degradasi yang menghasilkan fluktuasi sinyal yang bersifat acak atau tidak konsisten, seperti pada skenario *Gaussian noise*. Dalam kondisi tersebut, sinyal monitoring dapat berosilasi di sekitar ambang batas tertentu tanpa memberikan indikasi degradasi yang jelas dan berkelanjutan.

Keterbatasan-keterbatasan tersebut menunjukkan bahwa pendekatan single-metric memiliki ruang lingkup representasi yang terbatas dalam lingkungan produksi yang kompleks dan dinamis. Temuan ini menjadi dasar untuk mengevaluasi pendekatan monitoring alternatif yang mempertimbangkan berbagai dimensi degradasi model secara simultan.

## 4.5 Analisis Monitoring Berdasarkan Pendekatan Multi-Metric

Penentuan threshold composite health score bertujuan untuk memperoleh batas keputusan yang dapat digunakan secara operasional dalam mekanisme monitoring model pembelajaran mesin. Threshold ini berfungsi sebagai indikator transisi yang membedakan kondisi model yang masih dianggap sehat dengan kondisi model yang terindikasi mengalami degradasi dan memerlukan perhatian lebih lanjut.

Dalam penelitian ini, threshold tidak ditetapkan secara arbitrer atau berdasarkan nilai baku tertentu, melainkan diturunkan secara empiris melalui eksperimen terkontrol. Eksperimen dilakukan dengan mengamati perilaku composite health score secara temporal pada kondisi degradasi data yang bersifat gradual. Pendekatan ini dipilih untuk memastikan bahwa threshold yang dihasilkan mencerminkan dinamika nyata perubahan perilaku model di lingkungan produksi.

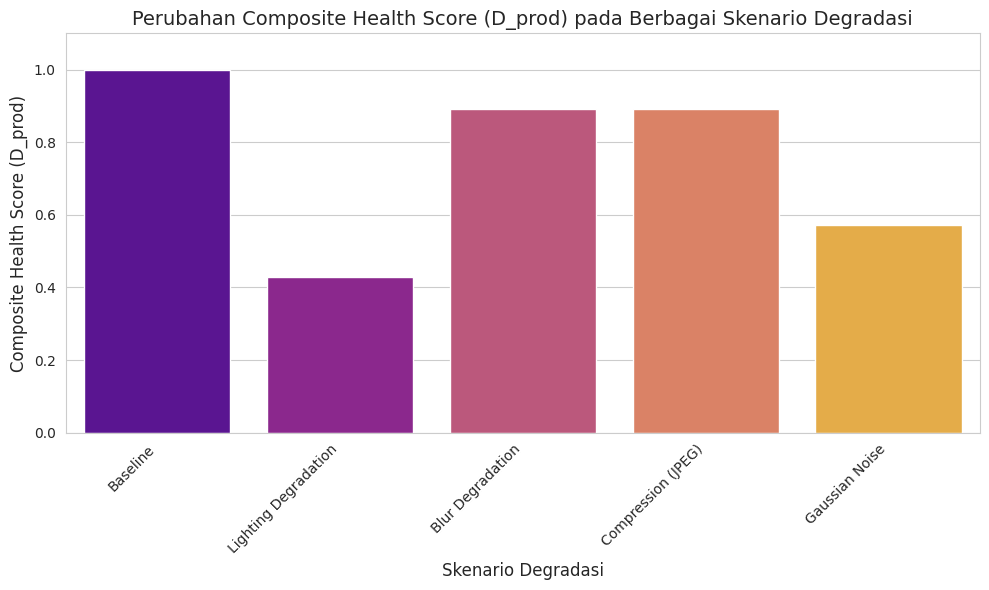
Sebagaimana telah dijelaskan pada Bab III, meskipun mekanisme monitoring dievaluasi pada lima kondisi skenario (baseline, blur, compression, lighting, dan Gaussian noise), eksperimen penentuan threshold dibatasi pada dua skenario degradasi, yaitu blur dan compression. Pembatasan ini dilakukan bukan untuk menghasilkan threshold yang spesifik terhadap skenario tertentu, melainkan untuk menyediakan dasar empiris yang stabil dan representatif dalam mengidentifikasi titik transisi composite health score dari kondisi sehat menuju terdegradasi.

Pemilihan skenario blur dan compression didasarkan pada karakteristik degradasinya yang bersifat gradual dan umum dijumpai pada sistem computer vision di lingkungan produksi. Kedua skenario tersebut memungkinkan observasi perubahan skor yang berlangsung secara bertahap dan konsisten, sehingga titik perubahan pola (change point) pada composite health score dapat diidentifikasi secara lebih jelas dibandingkan dengan degradasi yang bersifat sangat kontekstual atau stokastik.

Threshold ditentukan dengan mengamati rentang nilai composite health score pada kondisi baseline dan membandingkannya dengan pola penurunan skor yang muncul seiring meningkatnya tingkat degradasi. Batas keputusan dipilih pada titik di mana penurunan skor mulai menunjukkan tren yang berkelanjutan dan tidak lagi dapat dikategorikan sebagai fluktuasi normal. Dengan pendekatan ini, threshold merepresentasikan kondisi transisi yang relevan secara operasional, bukan sekadar perubahan statistik yang bersifat sesaat.

Setelah threshold ditetapkan, nilai tersebut digunakan sebagai acuan untuk menginterpretasikan hasil monitoring pada seluruh skenario degradasi yang dievaluasi dalam penelitian ini, termasuk lighting dan Gaussian noise. Dengan demikian, threshold diposisikan sebagai batas keputusan global untuk composite health score, bukan sebagai batas yang bergantung pada jenis degradasi tertentu.

Pendekatan ini memungkinkan composite health score dan threshold yang dihasilkan untuk digunakan secara konsisten dalam berbagai kondisi operasional. Meskipun nilai threshold bersifat kontekstual terhadap konfigurasi metrik dan data yang digunakan, mekanisme penentuannya memberikan kerangka kerja yang sistematis dan dapat direplikasi untuk kalibrasi threshold pada sistem dan domain yang berbeda



Gambar 4.5 Perubahan Composite Health Score (D\_prod) per Batch

Sumber: Penulis, 2025

4.5.1 Respons Multi-Metric terhadap Skenario Degradasi

Subbab ini membahas respons monitoring model berdasarkan pendekatan *multi-metric* terhadap berbagai skenario degradasi data. Pendekatan ini mengevaluasi kondisi model dengan mengamati beberapa metrik secara bersamaan, sehingga perubahan perilaku model dapat dianalisis secara lebih kontekstual.

Hasil eksperimen menunjukkan bahwa pendekatan multi-metric mampu menangkap perubahan sinyal monitoring secara lebih stabil dan konsisten dibandingkan pendekatan single-metric. Perubahan pada satu metrik dapat dikonfirmasi oleh metrik lainnya, sehingga interpretasi kondisi model tidak bergantung pada satu indikator tunggal.

Pada skenario degradasi yang bersifat gradual, seperti *blur*, sinyal multi-metric menunjukkan pola perubahan yang lebih terstruktur dan mudah diinterpretasikan. Sementara itu, pada skenario degradasi yang bersifat diskrit, seperti *compression (JPEG)*, pendekatan multi-metric mampu merefleksikan perubahan kondisi model secara lebih jelas melalui kombinasi beberapa indikator.

Selain itu, pada skenario degradasi yang menghasilkan fluktuasi sinyal, seperti *Gaussian noise*, pendekatan multi-metric menunjukkan kemampuan yang lebih baik dalam membedakan fluktuasi acak dari perubahan kondisi model yang bersifat konsisten dan berkelanjutan.

4.5.2 Keunggulan Pendekatan Multi-Metric

Keunggulan utama pendekatan multi-metric terletak pada kemampuannya dalam merepresentasikan kondisi kesehatan model secara lebih komprehensif. Dengan mengamati berbagai dimensi degradasi secara simultan, pendekatan ini mengurangi risiko interpretasi yang keliru akibat ketergantungan pada satu metrik saja.

Pendekatan multi-metric memungkinkan pemisahan yang lebih jelas antara perubahan distribusi data dan perubahan perilaku prediksi model. Informasi tambahan dari berbagai metrik membantu memberikan konteks terhadap sinyal monitoring yang muncul, sehingga kondisi model dapat dievaluasi dengan tingkat keyakinan yang lebih tinggi.

Selain itu, stabilitas sinyal yang dihasilkan oleh pendekatan multi-metric menjadikannya lebih sesuai untuk mendukung pengambilan keputusan operasional dalam lingkungan produksi. Representasi kondisi model yang lebih informatif dan konsisten memberikan landasan yang kuat untuk proses agregasi metrik ke dalam satu nilai komposit yang merepresentasikan tingkat kesehatan model secara keseluruhan.

Keunggulan-keunggulan tersebut menegaskan bahwa pendekatan multi-metric merupakan fondasi yang lebih tepat untuk membangun mekanisme *health check* model yang adaptif dan dapat diandalkan, sebagaimana direalisasikan melalui perhitungan *composite health score* pada subbab berikutnya.

## 4.6 Hasil Perhitungan Composite Health Score

Subbab ini menyajikan hasil perhitungan *composite health score* sebagai representasi agregat dari kondisi kesehatan model pada fase pasca-deployment. Nilai komposit ini diperoleh melalui proses penggabungan beberapa metrik monitoring yang telah dianalisis pada subbab sebelumnya, sehingga mencerminkan kondisi model secara menyeluruh dalam satu indikator terintegrasi.

Hasil perhitungan menunjukkan bahwa *composite health score* mampu merepresentasikan perubahan kondisi model secara lebih konsisten dibandingkan pengamatan terhadap metrik individual. Pada kondisi baseline, nilai komposit berada pada rentang yang stabil, mencerminkan keadaan model yang sehat. Ketika skenario degradasi diterapkan, perubahan nilai komposit terlihat mengikuti pola degradasi yang terjadi, baik yang bersifat gradual maupun diskrit.

Pada skenario degradasi yang bersifat gradual, seperti *blur*, perubahan *composite health score* terjadi secara bertahap seiring meningkatnya tingkat degradasi. Pola ini mencerminkan kemampuan indikator komposit dalam menangkap perubahan kecil namun konsisten yang mungkin sulit diinterpretasikan apabila hanya mengandalkan satu metrik. Sebaliknya, pada skenario degradasi yang bersifat lebih diskrit, seperti *compression (JPEG)*, perubahan nilai komposit menunjukkan pergeseran yang lebih tegas dalam rentang batch yang relatif singkat.

Selain itu, pada skenario degradasi yang menghasilkan fluktuasi sinyal, seperti *Gaussian noise*, *composite health score* menunjukkan perilaku yang lebih stabil dibandingkan metrik individual. Proses agregasi multi-metrik berperan dalam meredam fluktuasi acak pada satu indikator, sehingga perubahan yang bersifat berkelanjutan dapat diidentifikasi dengan lebih jelas.

Secara keseluruhan, hasil pada subbab ini menunjukkan bahwa *composite health score* memberikan representasi kondisi kesehatan model yang lebih informatif dan mudah diinterpretasikan. Indikator ini menjadi dasar yang kuat untuk menentukan status kesehatan model secara operasional, yang akan dibahas lebih lanjut pada subbab berikutnya.

## 4.7 Penentuan dan Validasi Status Kesehatan Model

Subbab ini membahas proses penentuan dan validasi status kesehatan model berdasarkan nilai composite health score yang telah diperoleh pada subbab sebelumnya. Penentuan status kesehatan model bertujuan untuk menerjemahkan nilai komposit yang bersifat numerik menjadi kategori kondisi model yang lebih mudah diinterpretasikan secara operasional.

Status kesehatan model ditentukan dengan mengelompokkan nilai composite health score ke dalam beberapa kategori kondisi, seperti healthy, degraded, dan critical. Proses pengelompokan ini dilakukan dengan menggunakan ambang batas (threshold) yang ditetapkan berdasarkan karakteristik nilai komposit pada kondisi baseline serta respons model terhadap skenario degradasi yang dianalisis. Pendekatan ini memungkinkan interpretasi status model yang konsisten dan relevan dengan konteks operasional.

Validasi terhadap status kesehatan model dilakukan dengan membandingkan hasil klasifikasi status dengan pola degradasi yang diamati pada masing-masing skenario. Pada kondisi baseline, status kesehatan model secara konsisten berada pada kategori healthy, mencerminkan stabilitas perilaku model ketika tidak terdapat degradasi data. Sebaliknya, pada skenario degradasi, perubahan status kesehatan model menunjukkan kesesuaian dengan tingkat dan karakter degradasi yang diterapkan.

Hasil validasi menunjukkan bahwa status kesehatan model yang diturunkan dari composite health score mampu merepresentasikan kondisi model secara lebih konsisten dibandingkan penilaian berbasis satu metrik. Perubahan status yang dihasilkan bersifat lebih stabil antar batch data dan tidak mudah terpengaruh oleh fluktuasi sinyal yang bersifat sementara.

Dengan adanya mekanisme penentuan dan validasi status kesehatan model ini, hasil monitoring dapat digunakan sebagai dasar pengambilan keputusan operasional pada lingkungan produksi. Status kesehatan model yang terdefinisi dengan jelas memungkinkan sistem monitoring untuk memberikan peringatan dini (early warning) serta mendukung tindakan lanjutan yang sesuai, seperti investigasi lebih lanjut atau penyesuaian strategi pemeliharaan model.

## 4.8 Perbandingan Pendekatan Single-Metric dan Multi-Metric

Subbab ini menyajikan perbandingan antara pendekatan single-metric dan multi-metric dalam mendeteksi degradasi model pada fase pasca-deployment. Perbandingan dilakukan untuk mengevaluasi perbedaan karakteristik sinyal monitoring, stabilitas indikator, serta relevansi hasil monitoring dalam mendukung pengambilan keputusan operasional.

Berdasarkan hasil eksperimen, pendekatan single-metric menunjukkan kemampuan dalam mendeteksi perubahan distribusi data pada beberapa skenario degradasi. Namun, efektivitas pendekatan ini sangat bergantung pada karakteristik metrik yang digunakan serta jenis degradasi yang terjadi. Pada skenario degradasi tertentu, perubahan sinyal single-metric dapat muncul terlambat atau bersifat ambigu, sehingga menyulitkan interpretasi kondisi kesehatan model secara menyeluruh.

Sebaliknya, pendekatan multi-metric memberikan gambaran kondisi model yang lebih komprehensif dengan mengamati berbagai dimensi degradasi secara simultan. Kombinasi beberapa metrik memungkinkan konfirmasi silang terhadap sinyal yang muncul, sehingga perubahan kondisi model dapat diinterpretasikan dengan tingkat keyakinan yang lebih tinggi. Pendekatan ini juga menunjukkan stabilitas sinyal yang lebih baik, terutama pada skenario degradasi yang menghasilkan fluktuasi indikator.

Dari sisi waktu deteksi degradasi, pendekatan multi-metric cenderung mampu mengidentifikasi perubahan kondisi model secara lebih konsisten pada berbagai skenario. Meskipun pendekatan single-metric dapat memberikan indikasi awal pada kondisi tertentu, pendekatan multi-metric menunjukkan keunggulan dalam menjaga keseimbangan antara sensitivitas deteksi dan stabilitas sinyal.

Secara keseluruhan, hasil eksperimen menunjukkan bahwa pendekatan monitoring berbasis multi-metric memberikan representasi kondisi kesehatan model yang lebih konsisten dibandingkan pendekatan single-metric. Pada berbagai skenario degradasi, single-metric monitoring cenderung menghasilkan sinyal yang ambigu atau terlambat, terutama ketika perubahan distribusi data tidak langsung berdampak pada tingkat kepercayaan prediksi. Sebaliknya, integrasi beberapa metrik dalam composite health score memungkinkan deteksi degradasi dilakukan secara lebih dini dan stabil, sehingga lebih relevan untuk digunakan sebagai dasar pengambilan keputusan operasional dalam lingkungan produksi.

## 4.9 Pembahasan dalam Konteks MLOps dan Monitoring Produksi

Dalam konteks penerapan Machine Learning Operations (MLOps), hasil penelitian ini menunjukkan bahwa pendekatan monitoring model tidak dapat diperlakukan sebagai proses evaluasi metrik semata, melainkan sebagai bagian dari mekanisme pengambilan keputusan operasional. Composite health score yang diusulkan dalam penelitian ini berperan sebagai penghubung antara sinyal teknis degradasi model dan tindakan operasional yang perlu diambil dalam lingkungan produksi, seperti peningkatan pemantauan, penyesuaian sistem, atau rollback model. Dengan demikian, monitoring model diposisikan sebagai komponen aktif dalam siklus hidup model pembelajaran mesin, bukan hanya sebagai alat observasi pasif.

Pendekatan single-metric yang umum digunakan, seperti Population Stability Index (PSI), terbukti memiliki keterbatasan ketika dihadapkan pada dinamika lingkungan produksi yang kompleks. Hasil eksperimen menunjukkan bahwa perubahan distribusi data yang terdeteksi secara statistik tidak selalu berkorelasi langsung dengan penurunan perilaku prediksi model. Kondisi ini berpotensi menghasilkan sinyal monitoring yang ambigu apabila digunakan sebagai satu-satunya dasar pengambilan keputusan. Dalam praktik MLOps, ambiguitas tersebut dapat berdampak pada keterlambatan intervensi atau munculnya false alarm yang tidak relevan secara operasional.

Sebaliknya, integrasi beberapa metrik monitoring dalam composite health score memungkinkan sistem monitoring untuk menangkap degradasi model dari berbagai dimensi secara simultan. Kombinasi metrik yang merepresentasikan perubahan distribusi data, stabilitas prediksi, dan performa sistem memberikan sinyal yang lebih stabil dan konsisten terhadap kondisi kesehatan model. Hal ini menjadi krusial dalam konteks MLOps, di mana keputusan operasional harus diambil dengan mempertimbangkan keseimbangan antara risiko degradasi model dan stabilitas sistem produksi.

Selain itu, penggunaan pendekatan multi-criteria decision model dalam monitoring model memberikan fleksibilitas dalam menyesuaikan sistem monitoring dengan kebijakan dan kebutuhan operasional organisasi. Penetapan bobot dan threshold yang bersifat policy-based memungkinkan sistem monitoring untuk beradaptasi terhadap konteks produksi yang berbeda tanpa harus bergantung pada proses optimasi ulang model secara terus-menerus. Pendekatan ini selaras dengan praktik industri, di mana keputusan terkait model sering kali dipengaruhi oleh toleransi risiko, kebutuhan layanan, dan keterbatasan sumber daya, bukan semata-mata oleh performa metrik individual.

Dengan demikian, hasil penelitian ini menegaskan bahwa monitoring model berbasis composite health score lebih sesuai untuk diterapkan dalam lingkungan produksi yang dinamis dibandingkan pendekatan single-metric. Pendekatan ini tidak hanya meningkatkan akurasi deteksi degradasi model, tetapi juga memperkuat keterkaitan antara evaluasi teknis model dan proses pengambilan keputusan dalam kerangka MLOps. Temuan ini mendukung pandangan bahwa sistem monitoring model yang efektif harus dirancang sebagai bagian integral dari tata kelola operasional model pembelajaran mesin, bukan sebagai komponen tambahan yang berdiri sendiri.

# BAB V

# KESIMPULAN DAN REKOMENDASI

## 5.1 Kesimpulan

Penelitian ini bertujuan untuk mengevaluasi efektivitas pendekatan monitoring model pembelajaran mesin pada fase pasca-deployment, khususnya dalam mendeteksi degradasi model klasifikasi berbasis *computer vision*. Fokus utama penelitian adalah membandingkan pendekatan *single-metric monitoring* dengan pendekatan *multi-metric monitoring* yang diagregasikan menggunakan *composite health score*.

Berdasarkan hasil eksperimen, pendekatan single-metric mampu mendeteksi perubahan distribusi data pada beberapa skenario degradasi. Namun, pendekatan ini memiliki keterbatasan dalam merepresentasikan kondisi kesehatan model secara menyeluruh, terutama pada skenario degradasi yang bersifat gradual atau menghasilkan fluktuasi sinyal yang tidak konsisten.

Sebaliknya, pendekatan multi-metric menunjukkan kemampuan yang lebih baik dalam menangkap perubahan perilaku model secara kontekstual. Dengan mengamati berbagai dimensi degradasi secara simultan, pendekatan ini menghasilkan sinyal monitoring yang lebih stabil dan informatif dibandingkan pengamatan terhadap satu metrik tunggal.

Hasil perhitungan *composite health score* menunjukkan bahwa indikator komposit mampu merepresentasikan kondisi kesehatan model secara lebih konsisten pada berbagai skenario degradasi. Nilai komposit ini memudahkan interpretasi kondisi model serta mendukung proses penentuan status kesehatan model secara operasional.

Secara keseluruhan, penelitian ini menyimpulkan bahwa pendekatan multi-metric dengan *composite health score* lebih efektif dan relevan untuk digunakan sebagai mekanisme *health check* model pembelajaran mesin pada lingkungan produksi yang dinamis, dibandingkan pendekatan monitoring berbasis satu metrik

.

## 5.2 Implikasi Penelitian

Hasil penelitian ini memiliki implikasi penting baik dari sisi teknis maupun operasional dalam penerapan sistem MLOps. Dari sisi teknis, pendekatan multi-metric memberikan kerangka monitoring yang lebih komprehensif untuk mengamati degradasi model, dengan mempertimbangkan berbagai dimensi perubahan data dan perilaku model secara bersamaan.

Dari sisi operasional, penggunaan *composite health score* memungkinkan penyederhanaan proses pengambilan keputusan dalam monitoring model produksi. Indikator komposit yang terintegrasi dapat digunakan sebagai dasar untuk pemicu peringatan dini (*early warning*), eskalasi investigasi, atau tindakan pemeliharaan model lainnya, tanpa harus menganalisis setiap metrik secara terpisah.

Selain itu, pendekatan monitoring yang diusulkan selaras dengan praktik MLOps dan *AI governance*, khususnya dalam mendukung observabilitas, auditabilitas, dan pengelolaan risiko model pembelajaran mesin. Dengan menyediakan representasi kondisi model yang lebih stabil dan mudah diinterpretasikan, pendekatan ini dapat meningkatkan keandalan sistem AI dalam lingkungan produksi.

Implikasi praktis dari penelitian ini menunjukkan bahwa mekanisme monitoring berbasis multi-metric dan *composite health score* dapat diintegrasikan ke dalam pipeline MLOps yang sudah ada, sehingga mendukung pengelolaan model secara berkelanjutan tanpa menambah kompleksitas operasional yang berlebihan.

## 5.3 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menginterpretasikan hasil yang diperoleh. Pertama, eksperimen dilakukan pada satu domain aplikasi computer vision dengan model klasifikasi berbasis CNN, sehingga generalisasi hasil ke domain atau arsitektur model lain belum dapat dipastikan.

Kedua, skenario degradasi data yang dianalisis terbatas pada degradasi visual sintetis tertentu, seperti blur, compression, lighting variation, dan Gaussian noise. Kondisi degradasi data di lingkungan produksi yang sesungguhnya dapat bersifat lebih kompleks dan tidak terkontrol sepenuhnya.

Ketiga, penentuan ambang batas (threshold) dan pembobotan metrik dalam perhitungan composite health score masih bersifat statis dan bergantung pada konfigurasi eksperimen yang digunakan. Pendekatan ini belum mempertimbangkan adaptasi bobot atau threshold secara dinamis seiring perubahan karakteristik data operasional.

Keterbatasan-keterbatasan tersebut menunjukkan bahwa hasil penelitian ini perlu dipahami dalam konteks ruang lingkup dan desain eksperimen yang telah ditetapkan.

## 5.4 Saran untuk Penelitian Selanjutnya

Berdasarkan keterbatasan yang telah diidentifikasi, penelitian selanjutnya dapat mengembangkan pendekatan monitoring ini dengan memperluas domain aplikasi dan jenis model pembelajaran mesin yang dievaluasi, sehingga validitas dan generalisasi hasil dapat ditingkatkan.

Selain itu, penelitian lanjutan dapat mengeksplorasi penggunaan skenario degradasi data yang lebih beragam dan realistis, termasuk degradasi yang bersifat non-sintetis atau terjadi secara bertahap dalam jangka waktu yang panjang. Hal ini akan memberikan gambaran yang lebih mendekati kondisi operasional sesungguhnya.

Pengembangan metode adaptasi bobot dan ambang batas secara dinamis juga merupakan arah penelitian yang menarik untuk meningkatkan fleksibilitas composite health score. Pendekatan ini berpotensi meningkatkan sensitivitas dan stabilitas monitoring model dalam lingkungan produksi yang terus berubah.

Terakhir, integrasi mekanisme monitoring berbasis multi-metric dengan sistem online monitoring dan automated response dalam pipeline MLOps dapat menjadi fokus penelitian selanjutnya, guna mendukung pengelolaan model pembelajaran mesin yang lebih responsif dan berkelanjutan.

# DAFTAR PUSTAKA

[1] U. Bhatt *et al.*, “Explainable machine learning in deployment,” *FAT\* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*, no. January, pp. 648–657, 2020, doi: 10.1145/3351095.3375624.

[2] A. K. Schalkamp *et al.*, “Machine learning-based personalized composite score dissects risk and protective factors for cognitive and motor function in older participants,” *Front. Aging Neurosci.*, vol. 16, no. October, pp. 1–11, 2024, doi: 10.3389/fnagi.2024.1447944.

[3] A. Kore *et al.*, “Empirical data drift detection experiments on real-world medical imaging data,” *Nat. Commun.*, vol. 15, no. 1, 2024, doi: 10.1038/s41467-024-46142-w.

[4] S. McFadden *et al.*, “DRMD: Deep Reinforcement Learning for Malware Detection under Concept Drift,” 2025, [Online]. Available: http://arxiv.org/abs/2508.18839

[5] S. Greco, B. Vacchetti, D. Apiletti, and T. Cerquitelli, “Unsupervised Concept Drift Detection From Deep Learning Representations in Real-Time,” *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 10, pp. 6232–6245, 2025, doi: 10.1109/TKDE.2025.3593123.

[6] O. Article, “Model Evaluation Beyond AUC: A Comparative Study of Somers’ D, Log Loss, Population Stability Index (PSI), and Kolmogorov–Smirnov (KS) Statistic in Credit Risk and Healthcare Prediction Models,” 2025, doi: 10.63282/3050-9246/icrtcsit-113.

[7] F. Bayram, B. S. Ahmed, and A. Kassler, “From concept drift to model degradation: An overview on performance-aware drift detectors,” *Knowledge-Based Syst.*, vol. 245, p. 108632, 2022, doi: 10.1016/j.knosys.2022.108632.

[8] D. Vela, A. Sharp, R. Zhang, T. Nguyen, and A. Hoang, “Temporal quality degradation in AI models,” *Sci. Rep.*, pp. 1–12, 2022, doi: 10.1038/s41598-022-15245-z.

[9] A. Malinin *et al.*, “Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks,” *Adv. Neural Inf. Process. Syst.*, no. NeurIPS, pp. 1–44, 2021.

[10] O. Mahdi, S. Bevinakoppa, and S. Singh, “Multi-Criteria Analysis of Concept Drift Detection Algorithms: A Decision-Making Approach,” *ICETAS 2024 - 9th IEEE Int. Conf. Eng. Technol. Appl. Sci.*, 2024, doi: 10.1109/ICETAS62372.2024.11119834.

[11] Y. Zhang, Q. Vera Liao, and R. K. E. Bellamy, “Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” *FAT\* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*, no. January, pp. 295–305, 2020, doi: 10.1145/3351095.3372852.

[12] L. Liu *et al.*, “Deep Learning for Generic Object Detection: A Survey,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020, doi: 10.1007/s11263-019-01247-4.

[13] V. Cortellessa and L. Traini, “Detecting latency degradation patterns in service-based systems,” *ICPE 2020 - Proc. ACM/SPEC Int. Conf. Perform. Eng.*, no. January, pp. 161–172, 2020, doi: 10.1145/3358960.3379126.

[14] H. Taherdoost and M. Madanchian, “Multi-Criteria Decision Making (MCDM) Methods and Concepts,” *Encyclopedia*, vol. 3, no. 1, pp. 77–87, 2023, doi: 10.3390/encyclopedia3010006.

[15] D. O. Hanchuk and S. O. Semerikov, “Implementing MLOps practices for effective machine learning model deployment: A meta synthesis,” *CEUR Workshop Proc.*, vol. 3918, pp. 329–337, 2025.

[16] A. Singla, “Machine Learning Operations (MLOps): Challenges and Strategies,” *J. Knowl. Learn. Sci. Technol. ISSN 2959-6386*, vol. 2, no. 3, pp. 333–340, 2023, doi: 10.60087/jklst.vol2.n3.p340.

[17] H. Tercan and T. Meisen, “Machine learning and deep learning based predictive quality in manufacturing: a systematic review,” *J. Intell. Manuf.*, vol. 33, no. 7, pp. 1879–1905, 2022, doi: 10.1007/s10845-022-01963-8.

[18] S. Studer *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, 2021, doi: 10.3390/make3020020.

[19] A. Paleyes, R. G. Urma, and N. D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” *ACM Comput. Surv.*, vol. 55, no. 6, 2023, doi: 10.1145/3533378.

[20] D. Kreuzberger, N. Kuhl, and S. Hirschl, “Machine Learning Operations (MLOps): Overview, Definition, and Architecture,” *IEEE Access*, vol. 11, no. February, pp. 31866–31879, 2023, doi: 10.1109/ACCESS.2023.3262138.

[21] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, “Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions,” *Sustain.*, vol. 12, no. 2, 2020, doi: 10.3390/su12020492.

[22] S. Agrahari and A. K. Singh, “Concept Drift Detection in Data Stream Mining: A literature review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9523–9540, 2022, doi: 10.1016/j.jksuci.2021.11.006.

[23] F. E. Casado, D. Lema, M. F. Criado, R. Iglesias, C. V. Regueiro, and S. Barro, “Concept drift detection and adaptation for federated and continual learning,” *Multimed. Tools Appl.*, vol. 81, no. 3, pp. 3397–3419, 2022, doi: 10.1007/s11042-021-11219-x.

[24] B. Eken, S. Pallewatta, N. Tran, A. Tosun, and M. A. Babar, “A Multivocal Review of MLOps Practices, Challenges and Open Issues,” *ACM Comput. Surv.*, vol. 58, no. 2, 2025, doi: 10.1145/3747346.

[25] J. F. Kurian and M. Allali, “Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications,” *J. Data, Inf. Manag.*, vol. 6, no. 3, pp. 207–216, 2024, doi: 10.1007/s42488-024-00119-y.

[26] J. du Pisanie and I. Visagie, “On testing the hypothesis of population stability for credit risk scorecards,” *ORiON*, vol. 36, no. 1, pp. 19–34, 2020, doi: 10.5784/36-1-678.

[27] Y. Chen, M. Jain, V. Sawhney, and L. Wu, “Enhancing Reliability in Recommendation Systems: Beyond point estimations to monitor population stability,” *CEUR Workshop Proc.*, vol. 3788, pp. 1–7, 2024.

[28] S. J. Warnett, E. Ntentos, and U. Zdun, “A model-driven, metrics-based approach to assessing support for quality aspects in MLOps system architectures,” *J. Syst. Softw.*, vol. 220, no. October 2024, p. 112257, 2025, doi: 10.1016/j.jss.2024.112257.

[29] B. Xia, Y. Liu, Q. Lu, L. Zhu, D. Sejdinovic, and X. Xu, “AgentOps Pattern Catalogue : Architectural Patterns for Safe and Observable Operations of Foundation Model-Based Agents”.

[30] C. Science and I. Technology, “Machine Learning – Driven Scaling Policies for Elastic Kubernetes Clusters,” 2023.

[31] S. Kavun1, “Multi-Metric Algorithmic Complexity: Beyond Asymptotic Analysis”.

[32] M. Baydaş and O. E. Elma, “An objective criteria proposal for the comparison of MCDM and weighting methods in financial performance measurement: An application in Borsa Istanbul,” *Decis. Mak. Appl. Manag. Eng.*, vol. 4, no. 2, pp. 257–279, 2021, doi: 10.31181/DMAME210402257B.

[33] R. Kumar *et al.*, “Multiple‐criteria decision‐making and sensitivity analysis for selection of materials for knee implant femoral component,” *Materials (Basel).*, vol. 14, no. 8, 2021, doi: 10.3390/ma14082084.

[34] R. Sharma *et al.*, “A Framework for Detecting Thyroid Cancer from Ultrasound and Histopathological Images Using Deep Learning, Meta-Heuristics, and MCDM Algorithms,” *J. Imaging*, vol. 9, no. 9, 2023, doi: 10.3390/jimaging9090173.

[35] L. N. Jescovitch *et al.*, “Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression,” *J. Sci. Educ. Technol.*, vol. 30, no. 2, pp. 150–167, 2021, doi: 10.1007/s10956-020-09858-0.

[36] M. González-Sierra *et al.*, “Relationship of Blood Inflammatory Composite Markers with Cardiovascular Risk Factors and Subclinical Atherosclerosis in Patients with Rheumatoid Arthritis,” *Life*, vol. 13, no. 7, pp. 1–15, 2023, doi: 10.3390/life13071469.

[37] M. Testi *et al.*, “MLOps : A Taxonomy and a Methodology,” *IEEE Access*, vol. 10, pp. 63606–63618, 2022, doi: 10.1109/ACCESS.2022.3181730.

[38] C. N. Wang, H. T. Tsai, T. P. Ho, V. T. Nguyen, and Y. F. Huang, “Multi-criteria decision making (MCDM) model for supplier evaluation and selection for oil production projects in vietnam,” *Processes*, vol. 8, no. 2, 2020, doi: 10.3390/pr8020134.

[39] I. Siksnelyte-Butkiene, E. K. Zavadskas, and D. Streimikiene, “Multi-criteria decision-making (MCDM) for the assessment of renewable energy technologies in a household: A review,” *Energies*, vol. 13, no. 5, 2020, doi: 10.3390/en13051164.

# **DAFTAR LAMPIRAN**

Lampiran 1 Daftar Riwayat Hidup

**DAFTAR RIWAYAT HIDUP**

1. **Biodata Mahasiswa**

NIM : 14230030

Nama Lengkap : Sony Harianto

Tempat dan Tanggal Lahir : Padang, 6 Maret 1991

Jenis Kelamin : Laki-Laki

Perumahan Bumi Cimanggis Indah 2, Kav. 11 Kel. Tapos, Kec. Tapos, Kota Depok, Jawa Barat

Alamat :

1. **Riwayat Pendidikan**
2. SDN 30 Cangkeh, Padang, Lulus Tahun 1997 - 2002
3. SMPN 11 Lubuk Kilangan, Padang, Lulus Tahun 2002 - 2005
4. SMKN 8 Lubuk Begalung, Padang, Lulus Tahun 2005 - 2008
5. STMIK Indonesia, Padang, Lulus Tahun 2010 - 2014
6. **Riwayat Pengalaman Pekerjaan**
7. IT Sys Admin, Tahun 2009 - 2014
8. Data Engineer, Tahun 2014-2017
9. Data Architect, Tahun 2017- Sekarang

A person in a suit with his arms crossed

AI-generated content may be incorrect.

Jakarta, 08 Januari 2026

Sony Harianto

Lampiran 2 Bukti Hasil Pengecekan Plagiarisme Laporan Tesis

A screenshot of a web page

AI-generated content may be incorrect.