2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

A Multi-Dimensional Taxonomy and Practical Framework for MLOps Monitoring

1st Tina Hafezi, 2nd Sara Pour Faraj, 3rd Seyed Mohammad Amin Alemohammad, *4th Niloofar Mirzaei Chahardeh

1st Dept. of Computer Engineering National University of Skills Tehran, Iran, Email: *TinaHafezii@gmail.com*

2nd Dept. of Computer Engineering National University of Skills Tehran, Iran, Email: *SarahPourfaraj@gmail.com*

3rd Dept. of Computer Engineering and Sciences CTB., Islamic Azad University, Tehran, Iran, Email: *MohammadAmin.Alemohammad@iau.ir*

*4th Dept. of Computer Engineering SR.C., Islamic Azad University Tehran, Iran, Email: *Niloofar.Mirzaeichahardeh@iau.ir* (Corresponding author)

Abstract

Unlike conventional software, ML systems operate in dynamic environments with non-stationary data, delayed labels, and changing behavior, making monitoring essential for reliability, stability, and long-term business value. This paper proposes a three-dimensional taxonomy of ML monitoring metrics covering data quality, statistical drift, classical performance indicators, behavioral proxies, system health, and business performance measures, including cost and value. Based on this taxonomy, it presents a practitioner-oriented framework for selecting, integrating, and operationalizing monitoring metrics across the ML lifecycle. The framework aligns metrics with stakeholder objectives, links technical and business indicators, and supports early-warning diagnostics, automated responses, label-dependent and label-independent monitoring, interpretability, human-in-the-loop evaluation, and automated retraining. It also defines observability modules, model lineage tracking, automation triggers.

The purpose of the research

- Machine learning models deployed in production environments face performance degradation due to evolving data distributions and changing real-world conditions [5].
- Delayed or unavailable ground-truth labels further complicate reliable model evaluation and monitoring processes.
- The study proposes a comprehensive MLOps monitoring framework for continuous assessment of machine learning systems.
- A structured framework is introduced to assist operational teams in selecting appropriate monitoring metrics and responding effectively to detected anomalies or failures [1].

Research method

1) MLOps in production:
   - MLOps evolved to move ML models from prototypes to production.
   - ML systems must handle non-stationary environments.
   - Data drift and concept drift make deployment more complex than traditional software.
2) Observability and Monitoring:
   - Monitoring relies on statistical observability metrics.
   - Distribution distance measures help detect data shifts.
   - Comprehensive observability includes:
     - data quality
     - model performance proxies
     - business metrics

Table I. METRICS TAXONOMY

Level	Subtype	Example Metrics	Purpose / Detects
Data-Level	Data Quality & Integrity	Completeness, Validity, Uniqueness	Detects input/pipeline errors
Data-Level	Schema Checks	Record Count, Cardinality	Detects structural changes
Data-Level	Drift Proxies	Moments, Quantiles, KS Distance, BC	Detects data drift without labels
Model-Level	Standard ML Performance	Accuracy, Precision, Recall, F1, AUC	Detects concept drift (needs labels)
Model-Level	Performance Stability (Forecasting)	Batch loss, Forecast loss	Detects instability, retraining needs
Model-Level	Output-Side / Behavioral Proxies	Class Distribution, Confidence Scores	Detects behavior drift (no labels yet)
Operational Level	System Metrics	Latency, Throughput, CPU/GPU, Memory	Detects system failures & inefficiency
Business-Level	Strategic Metrics	ROI, User Satisfaction	Measures overall business value
Business-Level	Cost Metrics	FN/FP Cost, CO_2 Cost, Labor Cost	Measures economic impact & risk

Practical or simulation results

- A practical framework is proposed for monitoring and maintaining machine learning systems in production environments.
- Continuous monitoring enables early detection of anomalies, performance degradation, and distribution shifts.
- Automated retraining mechanisms can be triggered when significant performance issues are identified.
- Human oversight remains essential to validate model.
- Effective MLOps practices improve long-term model stability, reduce operational costs, and enhance system reliability [6].
- Sustained model performance is critical for maintaining the practical value and effectiveness of machine learning applications.

conclusion

Monitoring ML model performance in production is a strategic imperative for AI system sustainability and reliability. This paper presented a structured taxonomy of MLOps monitoring metrics, categorized into data quality, model performance, operational health, and business impact [3][4].

Our framework emphasizes balancing automation and human oversight. While automation is essential for scaling operations and consistency, the complexity of concept drift often requires human-in-the-loop interpretation of ambiguous signals and high-stakes decisions. Adopting these strategies, prioritizing customer-centric metrics and integrating observability early, allows organizations to shift from reactive troubleshooting to proactive reliability [2][7].

References

[1] S. J. Warnett, E. Ntentos, and U. Zdun, “A model driven, metrics-based approach to assessing support for quality aspects in MLOps system architectures,” Journal of Systems and Software, vol. 220, p. 112257, Feb. 2025, doi: 10.1016/j.jss.2024.112257.

[2] S. Shankar and A. G. Parameswaran, “Towards Observability for Production Machine Learning Pipelines,” Proc. VLDB Endow., vol. 15, no. 13, pp. 4015–4022, Sept., doi: 10.14778/3565838.3565853. 2022

[3] P. Kourouklidis, D. Kolovos, J. Noppen, and N. Matragkas, “A Domain-Specific Language for Monitoring ML Model Performance,” in 2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), Oct. 2023, pp. 266–275.doi:10.1109/MODELSC59198.2023.00056

[4] D. Protschky, L. Lämmermann, P. Hofmann, and N. Urbach, “What Gets Measured Gets Improved: Monitoring Machine Learning Applications in Their Production Environments,” IEEE Access, vol. 13, pp. 34518–34538, 2025, 10.1109/ACCESS.2025.3534628.

[5] L. Berberi et al., “Machine learning operations landscape: platforms and tools,” Artif Intell Rev, vol. 58, no. 6, p. 167, Mar. 2025, doi: 10.1007/s10462025-11164-3.

[6] J. Leest, C. Raibulet, I. Gerostathopoulos, and P. Lago, “Expert Monitoring: Human-Centered Concept Drift Detection in Machine Learning Operations,” in 2024 IEEE/ACM 46th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), Apr. 2024, pp. 1–5. doi: 10.1145/3639476.3639771.

[7] A. P. Woźniak, M. Milczarek, and J. Woźniak, “MLOps Components, Tools, Process, and Metrics: A Systematic Literature Review,” IEEE Access, vol. 13, pp. 22166–22175, 10.1109/ACCESS.2025.3534990.