AEGIS: Toward Expert-in-the-loop Industrial Anomaly Detection

Dongmin Kim, Ye Seul Sim, Suhee Yoon, Sanghyu Yoon, Seungdong Yoa, Soonyoung Lee*, Woohyung Lim*

LG AI Research

{dmkim, ysl.sim, suhee.yoon, sanghyu.yoon, seungdong.yoa, soonyoung.lee, w.lim}@lgresearch.ai

Abstract

Anomaly detection platforms in real-world environments require continuous interaction between automated systems and domain experts, as anomalies evolve dynamically and their definitions vary across contexts. Therefore, an effective platform must collaborate with experts and incorporate their feedback to update the system. This paper introduces AEGIS, an anomaly detection platform that aims to support interaction between domain experts and data-driven agents through three core capabilities: (1) data-driven insights through real-time monitoring, explanations, and distribution shift detection, which invoke customized tools to generate appropriate responses, (2) an expert feedback interface for labeling and direct updates via chat-based interaction, and (3) autonomous model construction that leverages expert-labeled data with LLM-driven hyperparameter optimization. Through this design, AEGIS fosters continuous interaction in which the platform provides insights while experts guide model improvement, ensuring user intent is reflected and robustness is maintained under evolving data distributions.

Introduction

Industrial anomaly detection systems (Kharitonov et al. 2022; Alzarooni et al. 2025; Chandola, Banerjee, and Kumar 2009) require continuous adaptation to new datasets while maintaining reliable performance under evolving conditions. Interaction between experts and the system is essential, since anomalies are domain-dependent and expert knowledge is critical for accurate identification. However, traditional systems lack mechanisms for incorporating expert insight and adapting to changes, necessitating platforms that bridge the gap between automated detection and human expertise.

To address these challenges, we introduce AEGIS, an anomaly detection platform that integrates expert feedback, adaptive model updates, and real-time monitoring with explainability, aiming to improve models continuously while incorporating expert knowledge, as illustrated in Figure 1.

The key contribution of AEGIS lies in a unified framework that enables autonomous model building, integrates expert feedback through an interactive interface, provides explanations for anomalies, and supports continuous updates to maintain robustness under evolving data conditions.

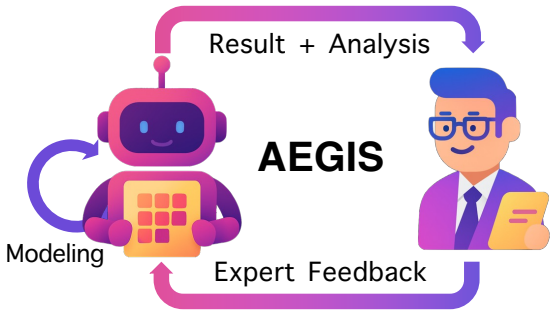

Figure 1: **Overview of AEGIS.** AEGIS builds models autonomously, and upon deployment, it provides results and analysis to experts. Experts can give feedback through the labeling interface to guide continuous model improvement.

System Architecture & Implementation

AEGIS is designed to support model building, expert interaction, monitoring, and adaptive updates as its core tasks, as illustrated in Figure 2. For this, LLM-based agents are implemented, utilizing customized tools from the Model Context Protocol (MCP) servers to generate responses.

Model Build

Inspired by recent advances in agentic model construction (Yang et al. 2025; Trirat, Jeong, and Hwang 2025), AEGIS adopts an evaluator–optimizer framework for building an anomaly detector. The build agent calls an evaluator tool in the MCP server with hyperparameters (e.g., learning rate, hidden dimensions), and the server returns performance results along with the trained models. Based on this feedback, the agent iteratively refines its hyperparameter search strategy. The base model is DeepSAD (Ruff et al. 2020), which primarily learns from normal data, while flexibly incorporating labeled anomalies. This design can be extended to other anomaly detection models (Livernoche et al. 2023; Ye et al. 2025; Chen et al. 2025) that support such settings.

System Monitoring with Chat Interface

The deployed model infers in real-time, and experts can interact with a chat-based interface. Anomaly scores are dis-

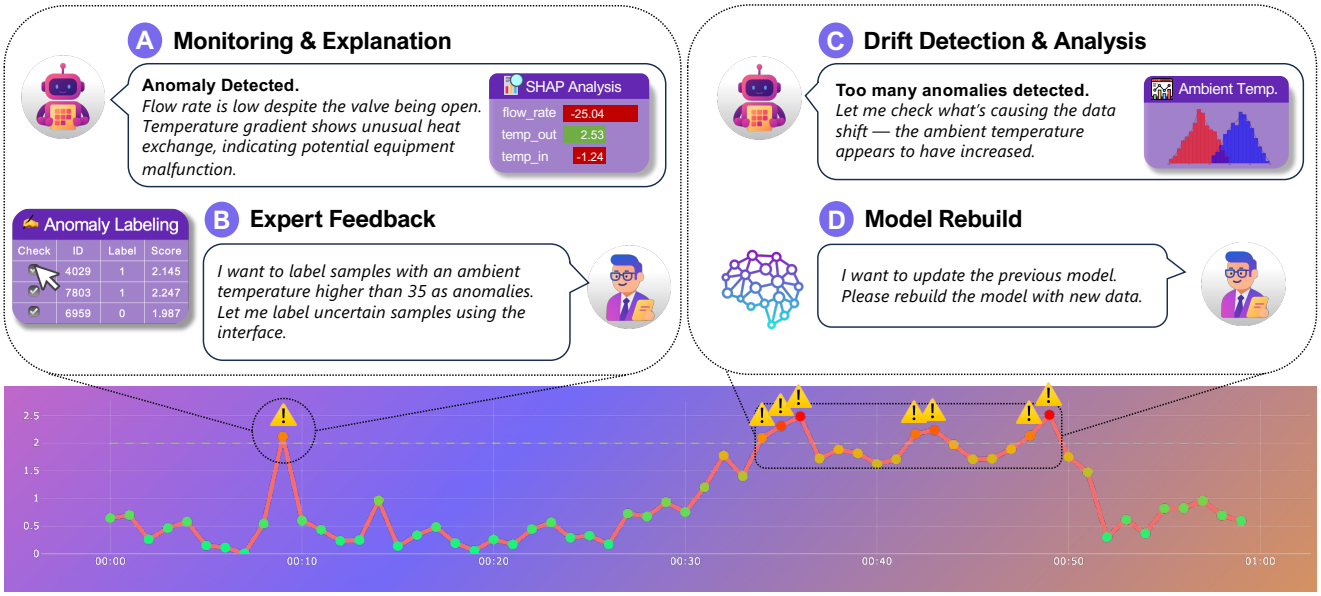

Figure 2: **AEGIS workflow according to system status.** AEGIS interacts with expert users and incorporates expert feedback into modeling. (A) The system issues alerts and generates explanations for anomalous samples. (B) Experts contribute feedback through the labeling interface and chat-based interaction. (C) Distribution shift alarms are raised, and status checks are supported with shift detection tools. (D) The agent rebuilds the model upon expert request, incorporating the new data and feedback.

played in real-time, and detected anomalies automatically generate alert messages appended to the chat history. When users request explanations for specific samples, AEGIS generates an answer based on the response from XAI tools returning SHAP attributes (Lundberg and Lee 2017), along with database queries for basic statistics and user-provided textual descriptions to give explanations.

User Feedback with Labeling Interface

Expert feedback is incorporated through a labeling interface that recommends high-uncertainty samples estimated using Monte Carlo Dropout (Gal and Ghahramani 2016), and through the chat interface specifying labeling rules on features. Upon such a request, AEGIS queries the database via MCP server to retrieve candidate samples and, upon confirmation, updates the labels through transaction calls. The labeled samples are incorporated into the next build phase, ensuring subsequent models integrate expert feedback.

Drift Detection and Rebuild

To ensure robustness under distribution shift, AEGIS monitors anomaly ratios and triggers alerts when they exceed predefined thresholds. Experts may then request a system status check, which invokes shift-measurement tools from Alibi Detect (Van Looveren et al. 2019), including spot-the-difference (Jitkrittum et al. 2016) and the Kolmogorov–Smirnov (K-S) test. Based on test results, AEGIS recommends whether to update the model, and upon a rebuild request, the agent re-enters the build phase, incorporating new labels and updated datasets to maintain performance under changing conditions.

Implementation Details

AEGIS utilizes custom tools provided by MCP servers when generating responses to users. Interaction with experts is mediated through a chat interface built on the Claude API (Anthropic 2025), which can further be extended to other LLM providers that support tool integration (Comanici et al. 2025; Fachada et al. 2025; Bai et al. 2023; Bae et al. 2025). The user interface of AEGIS is developed with Streamlit, offering an interactive environment for anomaly monitoring, labeling, and explanation. All data points and their corresponding metadata, including timestamps, labels, and uncertainty estimates, are stored in a MySQL database and connected to the agent through MCP tools that manipulate the database.

Conclusion and Future Work

AEGIS is designed with a primary focus on enabling interaction between the platform and industrial domain experts. AEGIS seeks to support reliable anomaly detection by integrating real-time monitoring and explanation, expert-guided feedback, and adaptive model updates through agent-driven construction. This feedback-driven design enables the system to adapt to evolving data distributions in real-world environments. For future work, AEGIS will extend its capabilities by incorporating text-based domain knowledge into modeling, developing causal analysis tools to provide insights into the origins of anomalies, and adopting advanced active learning methods to enhance the efficiency of expert labeling. Through these extensions, AEGIS is envisioned to evolve into a comprehensive anomaly detection platform that supports domain experts with interactions.

References

Alzarooni, A.; Iqbal, E.; Khan, S. U.; Javed, S.; Moyo, B.; and Abdulrahman, Y. 2025. Anomaly Detection for Industrial Applications, Its Challenges, Solutions, and Future Directions: A Review. *CoRR*, abs/2501.11310.

Anthropic. 2025. Claude Opus 4.1. https://www.anthropic.com/news/claude-opus-4-1. Accessed: 2025-08-20.

Bae, K.; Choi, E.; Choi, K.; Choi, S. J.; Choi, Y.; Han, K.; Hong, S.; Hwang, J.; Hwang, T.; et al. 2025. EXAONE 4.0: Unified Large Language Models Integrating Non-reasoning and Reasoning Modes. *arXiv preprint arXiv:2507.11407*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3): 15:1–15:58.

Chen, S.; Qian, Z.; Siu, W.; Hu, X.; Li, J.; Li, S.; Qin, Y.; Yang, T.; Xiao, Z.; Ye, W.; et al. 2025. Pyod 2: A python library for outlier detection with llm-powered model selection. In *Companion Proceedings of the ACM on Web Conference 2025*, 2807–2810.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fachada, N.; Fernandes, D.; Fernandes, C. M.; Ferreira-Saraiva, B. D.; and Matos-Carvalho, J. P. 2025. GPT-4.1 Sets the Standard in Automated Experiment Design Using Novel Python Libraries. *arXiv preprint arXiv:2508.00033*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. the International Conference on Machine Learning (ICML)*.

Jitkrittum, W.; Szabó, Z.; Chwialkowski, K. P.; and Gretton, A. 2016. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29.

Kharitonov, A.; Nahhas, A.; Pohl, M.; and Turowski, K. 2022. Comparative analysis of machine learning models for anomaly detection in manufacturing. *Procedia Computer Science*, 200: 1288–1297.

Livernoche, V.; Jain, V.; Hezaveh, Y.; and Ravanbakhsh, S. 2023. On diffusion modeling for anomaly detection. *arXiv preprint arXiv:2305.18593*.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.; and Kloft, M. 2020. Deep Semi-Supervised Anomaly Detection. In *Proc. the International Conference on Learning Representations (ICLR)*.

Trirat, P.; Jeong, W.; and Hwang, S. J. 2025. AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML. In *Forty-second International Conference on Machine Learning*.

Van Looveren, A.; Klaise, J.; Vacanti, G.; Cobb, O.; Scillitoe, A.; Samoilescu, R.; and Athorne, A. 2019. Alibi Detect: Algorithms for outlier, adversarial and drift detection.

Yang, T.; Liu, J.; Siu, W.; Wang, J.; Qian, Z.; Song, C.; Cheng, C.; Hu, X.; and Zhao, Y. 2025. AD-AGENT: A Multi-agent Framework for End-to-end Anomaly Detection. *arXiv preprint arXiv:2505.12594*.

Ye, H.; Zhao, H.; Fan, W.; Zhou, M.; dan Guo, D.; and Chang, Y. 2025. DRL: Decomposed Representation Learning for Tabular Anomaly Detection. In *The Thirteenth International Conference on Learning Representations*.