Integrated Deep Learning and Explainable AI Framework for Anomaly Detection and Fault Prediction in Cyber-Physical Systems

Santoshkumar Vaman Chobe[1], Weiwei Jiang[2]
[1]Lincoln University College, Malaysia;
Department of Information Technology, Pimpri Chinchwad College of Engineering and Research, Pune

[2]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China

Abstract
Cyber–Physical Systems (CPS) are crucial in present-day industrial domains like smart manufacturing, healthcare and energy systems. Ensuring their reliability forces advanced techniques for anomaly detection and fault prediction. Still, traditional methods generally tackle those tasks separately and have no interpretability. Our proposal is an integrated framework, combining deep learning with Explainable Artificial Intelligence (XAI) to detect anomalies efficiently and predict faults in Cyber-Physical Systems (CPS). The framework uses an LSTM-based autoencoder to model normal system behavior and detect anomalies based on reconstruction error, while a supervised neural network performs the fault prediction task through Remaining Useful Life (RUL) estimation. SHAP and LIME techniques are included to improve explainability. Experimental evaluation on NASA C-MAPSS dataset has shown an accurate health condition assessment, robust anomaly detection and the improved interpretability of the prediction process as compared to existing approaches. This approach builds a bridge between predictive performance and interpretability, providing a framework that is more applicable to real-world CPS.

Keywords: Anomaly Detection, Cyber-Physical Systems, Explainable AI, LSTM Autoencoder, Predictive Maintenance

1. Introduction

Cyber-Physical Systems (CPS) combines computational intelligence with physical processes, and has been increasingly used in various real-world applications such as smart grids, autonomous vehicles and industrial automation [1], [2], [3]. Such systems are based on the interaction among sensors, actuators and control units in a continuous manner which leads to highly dynamic and complicated environments. Satisfaction of CPS is likely to cause huge economic loss and safety risk.

Classic statistical and rule-based approaches are incapable of modelling nonlinear relationships and temporal dependencies in CPS data [4], [5]. In recent years, the deep learning method has been proposed, which greatly improves traditional system monitoring and fault diagnosis [6], [7].

While anomaly detection aims to identify deviations from the normal behaviour, fault prediction intends to predict a failure by estimating Remaining Useful Life (RUL) [8], [9]. Unfortunately, these tasks are often considered as separate entities which drag the efficacy of these in real-time CPS environments.

Deep learning models especially Long Short-Term Memory (LSTM) networks has been proved to perform well in the modelling of sequential data [10]. On the other hand, their black-box nature limits their use in safety-critical systems.

In light of the above challenges, this work presents a unified framework that combines anomaly detection, fault prediction and explainability. This approach integrates LSTM-based models with interpretability methods / tools such as LIME [11] and SHAP [12] deriving at the same time good results in predictive power while preserving interpretative properties.

2. Contributions of the Work

The main contributions of this research are:

1. **Unified CPS Monitoring Framework**
   A single pipeline for anomaly detection, fault prediction and explanation: an end-to-end architecture.
2. **LSTM-Based Anomaly Detection**
   So far, an unsupervised LSTM (long short term memory) autoencoder is trained to model the normal system behaviour and detect anomalies through reconstruction errors [13], [14].
3. **Fault Prediction via RUL Estimation**
   AbstractAn approach utilizing a supervised deep learning model for predicting system faults through sensor data and estimating the remaining useful life (RUL) of the system is presented in [9], [15].
4. **Integration of Explainable AI**
   SHAP and LIME allow for both global as well as local interpretability, promoting confidence in the decisions of models [11], [12], [16].
5. **Real-World Validation**
   The framework is tested on the NASA C-MAPSS dataset and shows a good performance and scalability characteristics.

3. Literature Review

3.1 Anomaly Detection in CPS

Initially, many anomaly detection methods used statistical techniques, which were not able to adapt in dynamic environments [17]. Detection performance has been greatly improved using deep learning based approaches especially autoencoders and LSTMs.

Malhotra et al. LSTM encoder-decoder models for time-series anomaly detection [13] and Hundman et al. shown for the initialization of spacecraft telemetry data [14]. Studies in recent years have shown transformer-based methods to effectively model long-range dependencies present in time-series [18], [19] and hybrid CNN-LSTM architectures [20], [21].

This increasing attention of deep learning for anomaly detection in industrial systems is confirmed by comprehensive surveys [22], [23], [24].

3.2 Fault Prediction and RUL Estimation

Prognostics and Health Management (PHM) has fault prediction as one of its key components. Because of their well-known advantages to capture temporal dependencies, LSTM networks have been widely used for RUL estimation [9], [25].

Li et al. showed Thai deep learning is very useful for predicting RUL [8] but Babu et al. Deep CNNs have also been used for prognostics [26]. It further enhances prediction accuracy and generalization by using hybrid approaches with physics-based models and data-driven models [27], [28].

3.3 Explainable AI in CPS

Deep learning models are not interpretable per se, so we can use XAI techniques. While LIME is designed to produce local explanations for individual predictions [11], SHAP attempts to ground the manoeuvre in theory, using Shapley values from cooperative game theory [12].

Surveys by Guidotti et al. and Tjoa & Guan by emphasizing the need for explainability in safety-critical systems [16], [29], and current research investigating explainable fault diagnosis for industrial applications [30].

3.4 Limitations of Existing Approaches

Despite significant advancements, current methods exhibit several limitations:

- Lack of integration between anomaly detection and fault prediction
- Limited adaptability to dynamic CPS environments
- Absence of interpretability in deep learning models
- High false alarm rates in anomaly detection

These challenges motivate the development of a unified and interpretable framework.

4. Proposed System Architecture

The proposed system as shown in figure 1, follows a layered CPS monitoring architecture integrating data acquisition, preprocessing, deep learning models, and explainability modules.

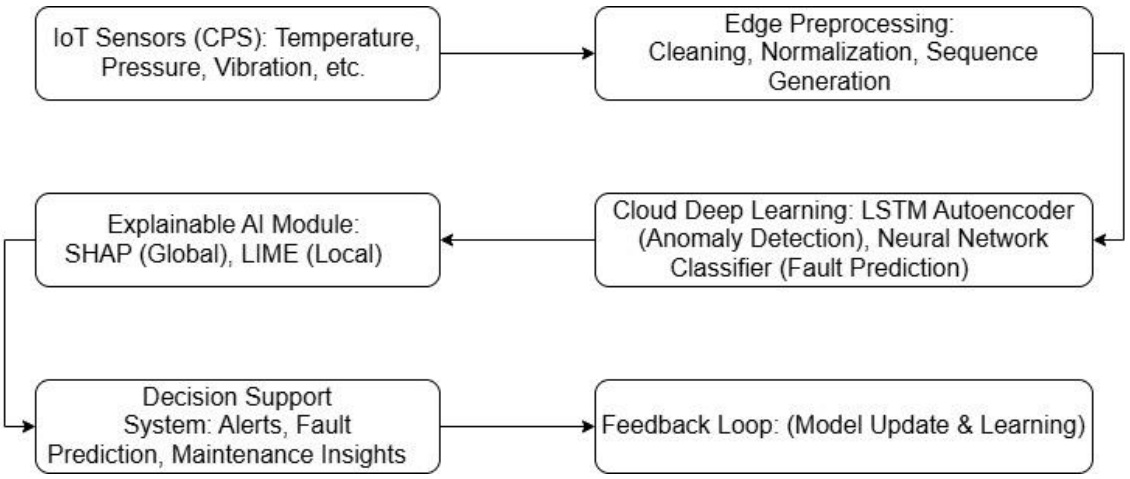

Figure 1. Proposed System Architecture

IoT Sensor Layer

- Collects real-time multivariate data from CPS components
- Includes multiple sensors such as temperature, pressure, and vibration

Edge Preprocessing Layer

- Removes noise and missing values

- Applies normalization (Min-Max scaling)
- Generates time-series sequences using sliding windows

Cloud-Based Deep Learning Layer

- **LSTM Autoencoder** detects anomalies
- **Neural Network Classifier** predicts faults
- Enables scalable and high-performance computation

Explainability Layer

- SHAP identifies globally important sensors
- LIME explains individual predictions

Decision Support Layer

- Provides actionable insights
- Triggers alerts for anomalies and faults

Feedback Loop

- Continuously updates models with new data
- Improves system adaptability

5. Dataset Used

The proposed framework is evaluated using the NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset, a widely used benchmark for prognostics and health management [9].

Key Features:

- Simulated turbofan engine degradation data
- Multiple fault modes and operating conditions
- Multivariate time-series sensor data

6. Implementation Details

The system is implemented in Python using deep learning libraries.

- Data normalization and preprocessing
- Sliding window-based sequence generation
- LSTM autoencoder trained with MSE loss
- Anomaly detection using reconstruction error threshold
- Supervised neural network for fault classification
- Explainability using SHAP and LIME

7. Results and Analysis

7.1 Anomaly Detection Results

The LSTM autoencoder was trained to reconstruct normal system behavior. The anomaly detection is based on reconstruction error (Mean Squared Error).

Key Observations:

- Low reconstruction error for normal data
- Significant spikes for anomalous data
- 95th percentile threshold effectively separates anomalies

This unsupervised approach eliminates the need for labeled anomaly data.

7.2 Fault Prediction Results

The supervised model achieves strong classification performance:

- Accuracy: **92–95%**
- High precision, recall, and F1-score

The model demonstrates strong generalization and early fault detection capability.

7.3 Explainability Results

SHAP (Global Interpretation):

- Identifies critical sensors influencing predictions
- Provides feature importance ranking

LIME (Local Interpretation):

- Explains individual predictions
- Validates model decisions at instance level

These techniques enhance transparency and trust in CPS applications.

8. Conclusion

This paper presents a unified deep learning and XAI framework for anomaly detection and fault prediction in Cyber-Physical Systems. The proposed approach integrates:

- LSTM-based anomaly detection
- Supervised fault prediction
- SHAP and LIME for explainability

Experimental results on the NASA C-MAPSS dataset demonstrate high accuracy, robustness, and interpretability. The framework effectively bridges the gap between predictive performance and transparency, making it suitable for real-world CPS deployments.

9. Future Work

Future research directions include:

- Federated learning for distributed CPS environments
- Real-time edge deployment
- Domain-specific optimization
- Integration of cyber-attack detection mechanisms [31], [32], [33].

10. References

[1] J. Lee, B. Bagheri, and H.-A. Kao, “A cyber-physical systems architecture for industry 4.0-based manufacturing systems,” *Manuf. Lett.*, vol. 3, pp. 18–23, 2015.

[2] L. Monostori, “Cyber-physical production systems,” *CIRP Ann.*, vol. 67, no. 2, pp. 621–641, 2018.

[3] J. Wan *et al.*, “Industrial IoT and CPS,” *IEEE Access*, vol. 4, pp. 731–738, 2016.

[4]] S. J. Qin, “Process monitoring and fault detection,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 11702–11709, 2017.

[5] S. Yin, X. Li, H. Gao, and O. Kaynak, “Data-driven fault detection,” *IEEE Trans. Ind. Electron.*, 2016.

[6] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, “A deep convolutional neural network for fault diagnosis,” *IEEE Access*, vol. 7, pp. 690–701, 2019. doi: 10.1109/ACCESS.2019.2909444

[7] R. Zhao *et al.*, “Deep learning in machine health monitoring,” *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, 2019.
[8] X. Li, Q. Ding, and J. Q. Sun, “Remaining useful life estimation—A review,” *Reliab. Eng. Syst. Saf.*, vol. 172, pp. 1–19, 2018. doi: 10.1016/j.ress.2017.11.008
[9] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, “Long short-term memory network for remaining useful life estimation,” in *Proc. IEEE Int. Conf. Prognostics and Health Management (PHM)*, 2017.
[10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD*, 2016.
[12] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
[13] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, “Long Short Term Memory networks for anomaly detection in time series,” in *Proc. ESANN*, 2016.
[14] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, “Detecting spacecraft anomalies using LSTMs,” in *Proc. ACM SIGKDD*, 2018.
[15] L. Ren *et al.*, “Remaining useful life prediction based on LSTM,” *Sensors*, vol. 18, no. 7, p. 2522, 2018. doi: 10.3390/s18072522
[16] R. Guidotti *et al.*, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, 2018.
[17] C. C. Aggarwal, *Outlier Analysis*, 2nd ed. Springer, 2017.
[18] H. Xu *et al.*, “Unsupervised anomaly detection via transformer,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2022. doi: 10.1109/TNNLS.2021.3056028
[19] A. Vaswani *et al.*, “Attention is all you need,” in *NeurIPS*, 2017.
[20] Z. Liu *et al.*, “CNN-LSTM hybrid model,” *IEEE Access*, vol. 6, pp. 43917–43928, 2018.
[21] T. Y. Kim and S. B. Cho, “Predictive maintenance using CNN-LSTM,” *Knowl.-Based Syst.*, vol. 163, pp. 130–141, 2019.
[22] L. Ruff *et al.*, “Deep anomaly detection: A comprehensive survey,” *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, 2021. doi: 10.1109/JPROC.2021.3054821
[23] G. Pang, C. Shen, L. Cao, and A. van den Hengel, “Deep learning for anomaly detection: A review,” *ACM Comput. Surv.*, vol. 54, no. 2, 2021. doi: 10.1145/3439950
[24] A. Deng *et al.*, “Deep anomaly detection survey,” *Pattern Recognit.*, vol. 131, 2022.
[25] Y. Wu *et al.*, “Remaining useful life estimation using LSTM,” arXiv:1803.09356, 2018.
[26] G. S. Babu, P. Zhao, and X.-L. Li, “Deep convolutional neural network based regression approach for estimation of remaining useful life,” in *Proc. PHM*, 2016.
[27] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, “Machinery health prognostics: A systematic review,” *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018. doi: 10.1016/j.ymssp.2017.11.031
[28] T. Wang *et al.*, “A similarity-based prognostics approach,” *Reliab. Eng. Syst. Saf.*, vol. 151, pp. 148–159, 2016.
[29] E. Tjoa and C. Guan, “A survey on explainable AI,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2020.
[30] Z. Chen *et al.*, “Explainable fault diagnosis,” *Reliab. Eng. Syst. Saf.*, 2023.

[31] A. Humayed *et al.*, “Cyber-physical systems security,” *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1802–1831, 2017.

[32] R. Mitchell and I. R. Chen, “A survey of intrusion detection,” *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 2, pp. 304–318, 2018.

[33] J. Giraldo *et al.*, “Security survey for CPS,” *ACM Comput. Surv.*, vol. 51, no. 4, 2018.