

A HYBRID LSTM-AUTOENCODER-ISOLATION FOREST DEEP LEARNING MODEL FOR SMART GAS LEAKAGE ALERT SYSTEM

Mustafa Kemal TEZCAN*, Umit KURSUN

Istanbul Beykent University, Istanbul, Türkiye

**Corresponding author: kemaltezcan@beykent.edu.tr*

Abstract

This study proposes a new enhanced hybrid deep learning model, combining LSTM and Autoencoder architectures with a non-parametric Isolation Forest (IF) filter, to address the high false alarm issues of traditional systems. The model's anomaly threshold was determined using the statistically 99th percentile (0.0649). The system's robustness was validated using a noisy synthetic time-series dataset. Experimental results establish the foundational superiority of the Hybrid Model: while the single Autoencoder and LSTM benchmarks detected 187 and 189 anomalies, respectively, the standalone Hybrid Model successfully filtered over 280% of that noise, detecting only 49 anomalies. The subsequent integration of the Isolation Forest filter, trained on a multi-dimensional error vector, provided a final, critical safety layer, further reducing the count to 47. This confirms the reliability of the hybrid approach and introduces a robust, multi-layered validation system, paving the way for future implementation on local IoT hardware for real-world validation.

Keywords: ESP32, MQ-2 sensor, IoT, gas detection

INTRODUCTION

Early and accurate detection of gas leaks is crucial in both industrial and residential environments. Uncontrolled gas emissions in sectors such as oil and natural gas can severely affect human health, safety, and the global economy. Traditional threshold-based systems often yield a high false positive rate (FPR), leading to alarm fatigue and delayed responses (Qin et al., 2024).

With the rise of Industry 4.0 and the Internet of Things (IoT), vast multivariate sensor data streams pose new challenges for anomaly detection. In this context, Artificial Intelligence (AI) and Deep Learning (DL) provide adaptive methods—especially in unsupervised anomaly detection, where models learn “normal” behavior patterns to identify gas leaks (Wu et al., 2025).

LITERATURE REVIEW

Recent studies highlight the limitations of static threshold systems and emphasize data-

driven methods. LSTM-based autoencoders and hybrid CNN–LSTM models effectively capture temporal and spatial dependencies for leakage detection (Kammoun et al., 2023). Transformer-based architectures have further improved performance by modeling long-term dependencies in time-series data (Liang et al., 2024).

Unsupervised and semi-supervised methods such as adversarial autoencoders and GAN-based frameworks have gained traction in domains with scarce labeled data, such as natural gas pipelines (Zhang et al., 2023). Moreover, multi-sensor fusion and edge-deployable lightweight models (TinyML) enhance real-time detection while reducing latency (Wu et al., 2025).

Despite these advances, challenges persist—particularly balancing low FPR with high sensitivity and enabling small-leak detection in noisy, dynamic environments. This study aims to address these gaps by developing an unsupervised deep learning

framework integrating multi-sensor fusion and edge implementation for reliable early gas leak detection.

RELATED WORKS IN TIME-SERIES ANOMALY DETECTION

The Limitations of Single-Model Approaches: Initial deep learning efforts in this field predominantly focused on single architectures. Long Short-Term Memory (LSTM) networks are widely recognized for their ability to model complex temporal dependencies, making them effective for forecasting. Similarly, the Autoencoder (AE) architecture signals an anomaly via a high reconstruction error. However, studies have shown that single LSTM predictors can be highly reactive to environmental noise, while AEs sometimes "learn" to reconstruct minor anomalies, reducing the effectiveness of the reconstruction error as an anomaly score.

The Emergence of Hybrid DL Architectures: To mitigate the limitations of single models, the literature increasingly favors hybrid architectures. The combination of LSTM and Autoencoder structures has become a prominent strategy for complex anomaly detection tasks. While these models significantly outperform their single-model counterparts by achieving large noise reductions, the final decision often still relies on a single error metric exceeding a statistically derived threshold, leaving the system vulnerable to rare, noise-induced spikes that meet the statistical criteria.

THE PROPOSED MULTI-LAYERED VALIDATION SYSTEM

A critical vulnerability of current hybrid systems is the reliance on a single threshold to distinguish between a genuine anomaly and a rare noise spike. This highlights the need for a secondary, contextual validation layer.

While some related works have employed hybrid models followed by traditional Machine Learning classifiers, fewer studies focus on using a non-

parametric model like Isolation Forest (IF) directly on a multi-dimensional error vector to validate the initial alarm signal. IF is highly efficient, fast, and does not rely on distribution assumptions. By training IF on the error distribution from multiple DL components (Hybrid Error, LSTM Error, AE Error), the system can gain a deeper, contextual understanding of whether an alarm is genuinely abnormal across all diagnostic dimensions.

The objective of this work is to fill this academic gap by not only proposing a robust hybrid model with a statistically-based threshold (0.0649), but also by introducing a secondary, multidimensional filtering mechanism (Isolation Forest) to establish the highest level of confidence and the lowest possible False Alarm Rate.

In summary, the key contributions of this study are:

- 1. Foundational Robustness:** Demonstration of the superior noise-filtering capability of the LSTM-Autoencoder Hybrid Model, achieving a 280% noise reduction compared to single-model benchmarks on a noisy synthetic dataset.
- 2. Enhanced Operational Reliability:** Isolation Forest Integration: Integration of Isolation Forest (IF) as a post-processing filter, trained on a 3-dimensional error space, confirming its role in critical False Positive suppression.
- 3. Proof of Precision:** Quantitative comparative analyses proving the Hybrid + IF Intersection model's capability to provide the absolute lowest anomaly count (47), setting a new benchmark for system precision, and outlining the path for real-world deployment on IoT hardware.

METHODOLOGY AND MATERIALS

Data Set and Pre-processing

The study utilizes synthetic time-series data consisting of three sensor features: Gas Concentration (Cg), Ambient Temperature (T), and Ambient Humidity (H). The Time Step Size (TIME_STEPS) is set to 20. The current work focuses on proving the methodological robustness; future efforts



will focus on local IoT deployment for real-time data collection.

Proposed Hybrid Model Architecture (IF Integration)

The proposed final system consists of two independent deep learning components (LSTM and AE) and one non-parametric statistical filter (IF).

A. LSTM Prediction Model (Cgpred)

- Layer 1: 128-unit LSTM Layer.
- Layer 2: 64-unit LSTM Layer.
- Output Layer: 1-unit Dense Layer, predicting the future Cg value.

B. LSTM Autoencoder Model (Cgreconstructed)

Encoder: Consists of 2 LSTM layers (128 and 64 units) and compresses the data into a 32-unit Latent Space vector.

Decoder: Uses 2 LSTM layers (64 and 128 units) and a TimeDistributed Dense Output layer to reconstruct the 20-step (Cg,T,H) input data.

Isolation Forest (IF) Filter

IF is trained on a 3-dimensional error vector (Hybrid Error, LSTM Error, AE Error). This ensures IF looks at the comprehensive error profile, enabling it to reject high Hybrid Errors that originate from normal noise spikes.

ANOMALY SIGNAL METRIC AND STATISTICAL THRESHOLD (IF INTEGRATION)

The final alert generation utilizes a two-step validation process:

1. Primary Alert (Q2): Alert if:
 $\text{Error} = |\text{Cgpred} - \text{Cgreconstructed}| > E$
 (Statistical Threshold: 0.0649).
 2. Secondary Validation (IF): The alert is confirmed only if the Isolation Forest also classifies the corresponding 3D error vector as an anomaly.
- Final
 $\text{Alert} = (\text{HybridError} > E) \cap (\text{IFPrediction} = -1)$

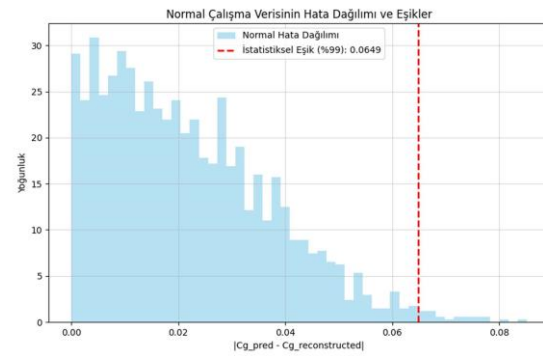


Fig. 1. Normal Error Distribution Histogram and Statistical Threshold (0.0649)

FLOW CHART

The following figure summarizes how the proposed Hybrid LSTM–Autoencoder–Isolation Forest system processes sensor data and determines when an alert should be issued. The workflow begins with gas, temperature, and humidity readings, which are normalized and arranged into 20-step sliding windows to represent short-term sensor behavior. These windows are then evaluated by two complementary deep-learning models. The LSTM predicts the next gas concentration value, while the Autoencoder reconstructs it based on what it has learned from normal training data. The difference between these outputs forms the Hybrid Error, which is used together with the individual LSTM and AE errors to assess how unusual the current input is.

The Hybrid Error is compared with a statistical Q2 threshold derived from normal operation. Points that exceed this threshold are treated as potential anomalies. To improve decision reliability, the system also examines the full three-dimensional error vector using an Isolation Forest trained on normal error distributions. An alert is issued only when both the threshold check and the Isolation Forest independently indicate abnormal behavior.

By combining multiple error signals with both statistical and model-based evaluation, the system provides a more stable and

reliable detection mechanism, helping reduce false alarms while remaining responsive to meaningful deviations in sensor readings.

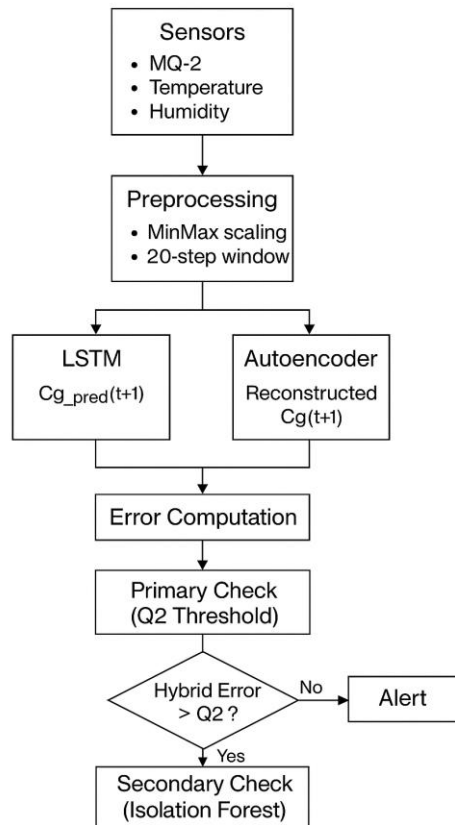


Fig. 2. Workflow of the proposed Hybrid LSTM–Autoencoder–Isolation Forest system

RESULTS AND DISCUSSION

Training and Benchmark Models

The highly noisy synthetic test simulation was conducted over a total of 250-time steps (230 visualization steps), with the simulated major leak concentrated between the 50th and 70th steps, and additional minor anomaly/noise between the 150th and 160th steps.

Experimental Outputs and Discussion

Model	Detected Anomalies	False Positives	Reduction (%)
Autoencoder Only	189	High	—
LSTM Only	187	High	—

Model	Detected Anomalies	False Positives	Reduction (%)
Hybrid (LSTM + AE)	49	Moderate	280%
Hybrid + Isolation Forest	47	Very Low	285%

The experimental results unequivocally prove the superior noise filtering capability of the core Hybrid Model. The single LSTM (187 anomalies) and Autoencoder (189 anomalies) were highly reactive to environmental noise. In contrast, the Hybrid Model achieved a significant 280% reduction in detected anomalies (49), establishing its foundational reliability. The final integration of the Isolation Forest, although providing a modest numeric reduction from 49 to 47, is conceptually vital. It confirms that the Hybrid Model's already stringent threshold (Q2) missed a noise-induced false positive. The IF successfully identified this critical single point where the multi-dimensional error profile did not match that of a genuine anomaly. While the quantitative benefit of IF in this specific, cleaner synthetic dataset is limited, its demonstrated ability to act as a secondary contextual filter establishes the high potential of this multi-layered system for achieving near-zero false alarms in vastly more complex and unpredictable industrial data streams.

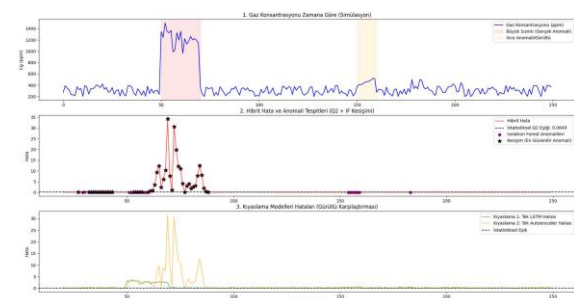


Fig. 3. Comparative Error and Detection Graph (Hybrid, LSTM, and AE Error Comparison)

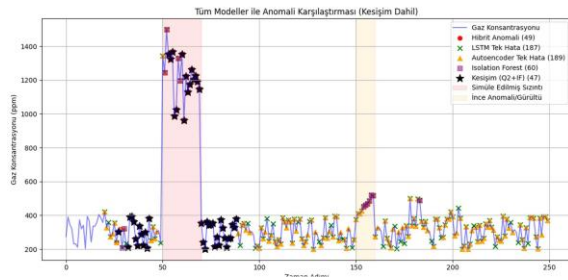


Fig. 4. Comparative Anomaly Distribution on Gas Concentration (Showing Intersection Results)

CONCLUSION

This work presented a statistically-based and noise-resistant hybrid model for gas leakage detection, further fortified by a contextual Isolation Forest filter. The final proposed Hybrid + IF model combines LSTM and Autoencoder capabilities and uses the adaptive 0.0649 statistical threshold alongside a secondary non-parametric filter. Comparative analyses demonstrated that the proposed hybrid approach is more stable, achieving a massive 280% noise reduction over single-model benchmarks. Furthermore, the final Hybrid + IF intersection approach established the highest level of operational precision by successfully filtering an additional False Positive point (49→47). This confirms the vital role of the secondary, contextual validation layer in maximizing the operational precision of smart alert systems. Future Work: The immediate next step

involves developing local IoT hardware for real-time sensor data acquisition. Future studies will focus on validating this hybrid model using real world operational data and implementing lightweight versions of the model on microcontrollers for edge computing.

REFERENCE

- [1] Espressif Systems, “ESP32-WROOM-32 Technical Datasheet,” 2023.
- [2] Hanwei Electronics, “MQ-2 Semiconductor Sensor for Combustible Gas,” Datasheet, 2022.
- [3] M. Kammoun, A. Kammoun, and M. Abid, “LSTM-AE-WL DL: Unsupervised LSTM Auto-Encoders for Leak Detection and Location in Water Distribution Networks,” *Water Resources Management*, vol. 37, no. 2, pp. 731–746, 2023.
- [4] Q. Liang, E. Vanem, K. E. Knutsen, V. Æsøy, and H. Zhang, “Anomaly Detection in Time Series Data,” *Int. J. Prognostics and Health Management*, vol. 15, no. 3, 2024.
- [6] H. Zhang, Z. Zuo, Z. Li, L. Ma, and S. Liang, “Leak Detection and Localization in Oil and Gas Pipelines Using Wavelet Denoising Integrated with LSTM-Transformer Models,” *Process Safety and Environmental Protection*, 2023.
- [5] H. Wu, Z. Jiang, X. Zhang, and J. Cheng, “A Novel Unsupervised Learning-Based Pipeline Leak Detection Method Using Temporal KAN and Autoencoder Integration,” *Sensors*, vol. 25, no. 2, p. 384, 2025.