Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P4806-cd

Trustworthy Anomaly Detection for Industrial Control Systems via Conformal Deep Autoencoder

Shuaiqi Yuan

Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, Special Administrative Region of China.

Xiaoge Zhang*

Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, Special Administrative Region of China. E-mail: xiaoge.zhang@polyu.edu.hk

Industrial control systems (ICSs) are critical infrastructures that remain highly vulnerable to both accidental and intentional anomalies, potentially leading to dangerous scenarios. While machine learning (ML) models are increasingly used for anomaly detection in ICSs, concerns about their trustworthiness persist due to their "black-box" nature, lack of effective uncertainty treatment, and absence of prediction guarantees. A key challenge is the high rate of false alarms, which can overwhelm operators and lead to unnecessary shutdowns. To address this, we propose a novel approach integrating deep autoencoders with conformal predictions to achieve high anomaly detection performance while providing statistical guarantees on false alarm rates. Our method uses conformal prediction as a post-hoc technique to enhance uncertainty treatment in a CNN-LSTM autoencoder, yielding trustworthy anomaly detection results with guaranteed false alarm rates. Recognizing temporal distribution shifts in time-series data, we incorporate temporal quantile adjustment to dynamically adapt the anomaly detection threshold, further improving temporal false alarm rate guarantees empirically. We validate the proposed model's ability to detect both accidental and attack-induced anomalies while maintaining a controlled false alarm rate using a publicly available dataset.

Keywords: Anomaly detection, Industrial control systems, Machine learning, Uncertainty treatment, Conformal predictions, Trustworthy AI, Process safety, Process security.

1. Introduction

With the increasing automation and digitization of industrial control systems (ICSs), safety and security issues remain a significant concern due to the potential for catastrophic consequences (Yuan et al., 2024). Anomaly detection plays a pivotal role in ensuring the safety and security of ICSs by identifying anomalies/outliers caused by either critical safety faults or malicious attacks. Consequently, various machine learning (ML) algorithms and models have been developed and applied to anomaly detection in ICSs, which are often characterized by operating multivariate time-series data.

Various ML and deep learning (DL) models have been applied for anomaly detection in an unsupervised manner, including Support Vector Machines (SVMs) (Anton et al., 2019), Random Forests (RF) (Alhaidari and Ezaz, 2019), Convolutional Neural Networks (CNNs) (Kravchik et al., 2018), Long Short-Term Memory (LSTM) networks (Perales et al., 2020), Generative Adversarial Networks (GANs) (Li et al., 2019), Graph Neural Networks (GNN) (Wu et al., 2021), Transformers (Shang et al., 2024), and autoencoders (Zhang et al., 2021), as well as hybrid approaches combining multiple models. With the advancement of DL, it is increasingly implemented for anomaly detection for ICSs due to its accuracy and capability to accommodate multivariate time-series data. Typically, DL-based anomaly detectors identify outliers based on the prediction errors, reconstruction errors of ML models, or a combination of both metrics. Among those DL models, autoencoder and its variants have been prevalent due to its high accuracy and capability to handle high-dimensional data and flexibility to integrate with other techniques.

However, concerns regarding the trustworthiness of ML-based anomaly detectors persist, primarily due to the black-box nature of ML models and the lack of guaranteed reliability in their results. A particularly pressing issue is the high incidence of false alarms (false positives) generated by anomaly detectors, which not only undermines their trustworthiness but also overwhelms operators in practical settings (Yang et al., 2024).

To address these challenges, ongoing advancements in uncertainty quantification (UQ) for ML models show promise in enhancing their suitability for ICS anomaly detection. UQ techniques can account for both epistemic and aleatory uncertainties in ML models, providing insights into prediction confidence and thereby improving trustworthiness. State-of-the-art UQ approaches include Bayesian neural networks, Gaussian process regression, Monte Carlo dropout, ensemble techniques, and hybrid methods (Nemani et al., 2023).

More recently, conformal prediction has gained significant attention in this domain due to its unique advantages. It provides statistical guarantees, distribution-free validity, and compatibility with various ML models in a wrap-up manner, all while maintaining computational efficiency (Angelopoulos & Bates, 2021). Given the demands for real-time anomaly detection, the ability to handle high-dimensional data, and the need to guarantee false alarm rates, the conformal prediction framework offers distinct benefits, making it a strong candidate for ICS applications.

To this end, this study integrates deep learning, specifically a deep autoencoder, with the conformal prediction framework to achieve trustworthy anomaly detection in ICS applications while ensuring guaranteed false alarm rates. The proposed model is designed to deliver reliable anomaly detection, enabling effective emergency responses to critical safety faults, failures, or malicious attacks. A publicly available ICS anomaly dataset is utilized to validate the performance of the conformal deep autoencoder in terms of its accuracy and false alarm rate guarantees.

2. Preliminaries

Preliminaries on deep autoencoders and conformal predictions are given below.

2.1. *Autoencoders for anomaly detection*

An autoencoder is a type of artificial neural network (ANN) designed to learn efficient representations of data, usually in an unsupervised manner (Kumar et al., 2024). It consists of two main components: encoder and decoder, which aim to compresses input data into a lower-dimensional latent representation and reconstruct the original input data from the latent representation, respectively.

In anomaly detection applications, autoencoder is trained on representative normal data to minimize the reconstruction error, which measures the difference between the original input and its reconstruction. When an anomaly instance is input into the autoencoder, the model struggles to reconstruct it accurately, leading to a higher reconstruction error. By thresholding reconstruction error properly, anomalies can be detected because anomalous data samples tend to have high reconstruction errors.

2.2. *Conformal predictions*

Conformal prediction (CP) is a user-friendly paradigm that provides statistically rigorous uncertainty quantification (Vovk et al., 2005). Conformal prediction works by calibrating model outputs to meet a desired confidence level. A key advantage of CP is its distribution-free validity, relying only on the assumption of data exchangeability. Combined with its post-hoc nature, CP is compatible with any machine learning model, making it particularly appealing for integration into complex deep learning systems (Angelopoulos & Bates, 2021).

CP methods can be categorized into full conformal prediction and split/inductive conformal prediction. Split/inductive CP significantly reduces computational costs while maintaining statistical validity, making it increasingly popular in practical applications. In this study, CP refers to split conformal prediction by default.

In the context of anomaly detection, CP demonstrates significant potential, especially due to its effectiveness in out-of-distribution detection. By offering statistical guarantees, CP enhances the trustworthiness of anomaly detection models and helps address the issue of false positives — a common challenge in ICS anomaly detection.

3. Methodology

3.1. *CNN-LSTM autoencoder*

Multivariate time-series data poses significant challenges for anomaly detection in ICS applications due to the difficulty of capturing complex spatial and temporal features. To address this, a CNN-LSTM autoencoder is employed. The CNN layers extract spatial correlations and dependencies within the multivariate data, while the LSTM layers capture temporal dynamics from the CNN outputs. This combination allows the model to effectively process the complex spatial-temporal features inherent in multivariate time-series data. Details of the CNN-LSTM autoencoder architecture are provided in Section 4.2.

3.2. *Conformal anomaly detection*

The use of conformal predictions for ML-based anomaly detection typically involves three steps: first, training a machine learning model on a training dataset; second, using a separate calibration dataset to compute nonconformity scores that quantify how unusual each data sample is; and finally, setting a nonconformity score threshold based on the calibration dataset with the desired confidence level and detecting anomalies in new data instances by comparing their nonconformity scores to this threshold.

In this study, the reconstruction error (mean squared error, MSE) of the autoencoder is used as the nonconformity score.

$$\mathrm{s}(X_i) = \frac{1}{n} \sum_{j=1}^{n} \left(x_{i,j} - \hat{x}_{i,j}\right)^2 \tag{1}$$

where nonconformity score $\mathrm{s}(X_i)$ is the MSE for data sample X_i . To achieve false positive guarantees, we make sure false positive rate less than a significance level α (e.g., $\alpha = 0.05$) (Angelopoulos & Bates, 2021).

$$P(C(X_{inlier}) = outlier) \leq \alpha \tag{2}$$

where C is the function used to detect outliers/anomalies, and X_{inlier} represents any new data instance from the normal dataset. To define the function C , the quantile of nonconformity scores from the calibration set is queried.

$$\hat{q} = quantile\left(s_1, \dots, s_n;\ \frac{[(n+1)(1-\alpha)]}{n}\right) \tag{3}$$

$$C(X) = \begin{cases} inlier & if\ s(X) \leq \hat{q} \\ outlier & if\ s(X) > \hat{q} \end{cases} \tag{4}$$

where s_1 to s_n represent the nonconformity scores of the data instances in the calibration set, and n is the size of the calibration set. $\hat{q}$ denotes the queried quantile of the nonconformity scores, and $\mathrm{s}(X)$ is the nonconformity score of a new data instance, X.

Additionally, P-value in the context of conformal prediction, indicating the proportion of calibration scores greater than or equal to a test sample's score, can also be leveraged to detect anomalies and is equivalent to method above. For a given test instance X, the p-value ($p(X)$) is calculated as:

$$p(X) = \frac{1+\sum_{i=1}^{n} 1\{s_i \geq s(X)\}}{n+1} \tag{5}$$

$$C(X) = \begin{cases} inlier & if\ p(x) \geq \alpha \\ outlier & if\ p(X) < \alpha \end{cases} \tag{6}$$

where $1\{s_i \geq s(X)\}$ is an indicator function that equals 1 if $s_i \geq s(X)$ and 0 otherwise.

3.3. *Temporal quantile adjustment*

Data exchangeability is a fundamental assumption in conformal predictions. Given time-series data, this assumption can hardly be met, leading to the loss of longitudinal/temporal coverage due to potential distribution shifts. As a result, this study implements a temporal quantile adjustment approach (TQA-B) proposed by Lin et al. (2022) to improve the temporal coverage empirically. The core idea of TQA-B is to adjust the quantile using δ, such that $\hat{\alpha} = \alpha - \delta$, to ensure the desired significance level is achieved. $\hat{\delta}$ is regarded as the prediction of δ and determined by a mapping function: $\hat{\delta}_t \leftarrow g(\hat{r}_t; \alpha)$. $\hat{r}_t$ is supposed to predict r_t, which is the rank of nonconformity scores in the calibration set at time t.

$$\bar{\varepsilon}_t = \sum_{t'=1}^{t} \frac{\beta^{(t-t')} \cdot s_{t'}}{t} \tag{7}$$

$$\hat{r}_t = Q^{-1}(\bar{\varepsilon}_t; \{\bar{\varepsilon}_t\}) \tag{8}$$

where $\bar{\varepsilon}_t$ represents the exponentially weighted nonconformity score at time t . β is a decay factor, set to 0.8 in this study. $s_{t'}$ denotes the nonconformity score at time t'. $\hat{r}_t$ is the quantile of $\bar{\varepsilon}_t$ in the set of weighted nonconformity scores, $\{\bar{\varepsilon}_t\}$. Q^{-1} is the inverse quantile function, which outputs the quantile of $\bar{\varepsilon}_t$ in $\{\bar{\varepsilon}\}$. Then, the adjusted quantile, $\hat{\delta}_t$, is calculated below.

$$\hat{\delta}_{t+1} = \begin{cases} Q \cdot \left(\hat{r}_t - (1-\alpha)\right) & if\ \hat{r}_t < 1-\alpha \\ \hat{r}_t - (1-\alpha) & if\ \hat{r}_t \geq 1-\alpha \end{cases} \tag{9}$$

$$Q = \frac{(2\alpha - \lfloor\alpha\rfloor)\cdot(\lfloor\alpha\rfloor+1)}{\lceil 1-\alpha\rceil\cdot\left((1-2\alpha)+1+\lfloor\alpha\rfloor\right)} \tag{10}$$

Where Q is a constant, and $\widehat{\alpha} = \alpha - \hat{\delta}$ is used to query the quantile in Eq. (3), instead of using α. Since $\widehat{\alpha}$ tends to approach 0, lead to overly conservative guarantees on the false alarm rate, bounding away $\widehat{\alpha}$ from 0 has been shown to empirically improve anomaly detection performance. To achieve this, $\delta = \lambda\hat{\delta}$ can be implemented (Lin et al., 2022). For instance, if we restrict $\widehat{\alpha} \geq 0.02$, we have:

$$\widehat{\alpha} = \alpha - \lambda\hat{\delta} \geq 0.02 \tag{11}$$

Considering Eq. (9) and (11), and given $\alpha = 0.05$, we then have $\lambda = \frac{\alpha - 0.02}{\alpha} = 0.6$.

To maintain a stable calibration set size, a sliding time window mechanism is employed. Since the calibration set is ideally intended to contain only in-distribution (normal) data, selection criteria are established to exclude obviously anomalous data, while ensuring the inclusion of in-distribution data.

$$T = quantile(S_{initial};\ \gamma) * (1+\mu) \tag{12}$$

$$\begin{cases} S_{t+1} = S_t(s_2, \ldots, s_t) \cup s(X) & s(X) < T \\ S_{t+1} = S_t & s(X) \geq T \end{cases} \tag{13}$$

Here, T represents the threshold for the nonconformity score. This threshold depends on quantile γ of the nonconformity scores from the initial calibration set ($S_{initial}$) and an inflation factor, μ. This operation helps to achieve a better false positive rate guarantee by ensuring the inclusion of in-distribution data, capturing distribution-shift features, and allowing for the inclusion of slightly anomalous data, which may raise the anomaly detection threshold slightly. In this study, γ is set to 1 and μ is configured to 0.01 to ensure that in-distribution data from the test set is incorporated. When the nonconformity score of a new test instance, $s(X)$, falls below the threshold T, calibration set's nonconformity score set slides by incorporating the new test instance and removing the oldest one in the set. Otherwise, the nonconformity score set remains unchanged, as shown in Eq. (13).

3.4. *Evaluation metrics*

To evaluate the performance of the conformal deep autoencoder in detecting anomalies and guaranteeing false positives, widely-used metrics for binary classification are used, including Precision, Recall, F1-Score, and AUROC (Area Under the Receiver Operating Characteristic Curve).

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{15}$$

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{16}$$

where TP represents true positives, which are correctly identified anomalies, FP denotes false positives, which are normal instances misclassified as anomalies, and FN refers to false negatives, which are anomalies misclassified as normal.

The AUROC evaluates the performance of the anomaly detector by capturing the trade-off between the True Positive Rate (Recall) and the False Positive Rate (FPR), the latter also known as the false alarm rate. An AUROC value of 1 signifies perfect discrimination, while a value of 0.5 indicates random guessing. False Positive Rate is calculated below.

$$\text{FPR} = \frac{FP}{FP+TN} \tag{17}$$

where TN represents true negatives, referring to normal instances correctly classified as normal.

4. Computational experiment and results

4.1. *Dataset descriptions*

We use a publicly available ICS dataset (Laso et al., 2017) for our numerical experiments. This dataset is well-suited for evaluating anomaly detection performance, as it encompasses a diverse range of anomalous patterns across 14 scenarios, including physical sabotage, system failures, and cyberattacks, in addition to normal operating conditions.

Given the multivariate time-series data, a time step of 0.05 seconds is implemented to sample the data. All 10 features, for both normal and anomalous conditions, are normalized to the range [0, 1]. To capture the inherent temporal dependencies, a sliding window of 50 time steps is employed, continuously generating data sequence samples.

Using unsupervised learning, the model is trained exclusively on normal data with a training set, taking 60% of the normal data. A calibration set taking 30% of the normal data is used to derive nonconformity scores and for conformal predictions. The test set, which includes the rest of the normal data and anomalous data, is employed to evaluate the model's performance. Details on the sizes of the training set, calibration set, and test set are provided in Table 1.

Table 1. Details on the split of the data sets.

Names of data sets	Data categories	Numbers of data samples
Training set	normal	81,660
Calibration set	normal	40,830
Test set	normal	13,611
	anomalous	6,453

4.2. *Model training*

A CNN-LSTM autoencoder, with the architecture shown in Fig. 1, is trained using the training set. This architecture captures both spatial and temporal features of the time-series data. The Mean Squared Error (MSE) is used as the loss function to minimize reconstruction errors, while the Adam optimizer is used for model training. Hyperparameters, including learning rate, batch size, and the number of epochs, are tuned to optimize the model's performance. The trained CNN-LSTM autoencoder then serves as the basis for conformal anomaly detection. Reconstruction errors generated by the autoencoder are used as nonconformity scores. A significance level, $\alpha = 0.05$, is applied, and a sliding calibration set is implemented to dynamically determine the anomaly detection threshold, according to the methods detailed in Sections 3.2 and 3.3.

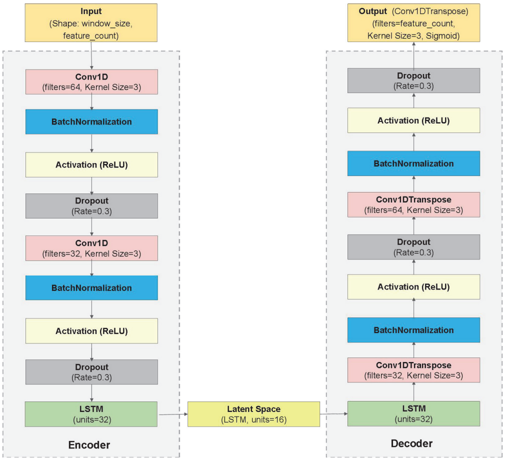

Fig. 1. Structure of the CNN-LSTM autoencoder.

4.3. *Results and analysis*

To evaluate the proposed approach from two perspectives — its anomaly detection performance and its ability to guarantee desired false alarm rates — we compared it against several baseline methods. These include Principal Component Analysis (PCA) with Hotelling's T^2 and Q-statistics for anomaly detection, k-Nearest Neighbors (k-NN), One-Class Support Vector Machine (OC-SVM), Isolation Forest, and a hybrid CNN-LSTM model. To ensure a fair comparison, all methods were trained and tested on the same datasets. Additional details on the baseline methods are provided below.

In the PCA-based approach, the anomaly detection indices — Hotelling's T^2 and Q-statistics (or Squared Prediction Error, SPE) — are used together to threshold anomalies, following the methodology outlined in Hashim et al. (2020). A 95% confidence level is applied to determine the threshold. This threshold theoretically controls false alarm rates under specific assumptions, such as linearly correlated Gaussian data, making this approach a suitable benchmark for evaluating the false alarm rate guarantees of our method. Furthermore, the number of retained PCA components is optimized to maximize the F1-score for each detection scheme, ensuring their best possible performance.

It is important to note that basic k-NN, OC-SVM, and Isolation Forest are not inherently designed to handle temporal dynamics. To address this limitation, a sliding window operation has been implemented to extract sequences from the time series, as described in Section 4.1. Flattening the raw window values into a vector before applying these methods for anomaly detection improves their capacity to capture temporal patterns.

In the k-NN scheme, a distance or similarity metric plays a critical role in anomaly detection. Euclidean distance is employed as the distance metric due to its widespread use for continuous attributes and its computational efficiency (Zhao et al., 2018). The 95th percentile of the Euclidean distance of the training set is utilized as the threshold for identifying anomalies, while the number of neighbors (k) is fine-tuned to maximize the F1-score. Similarly, in the OC-SVM scheme, a baseline One-Class SVM from the scikit-learn library (Scikit, 2025) is used for anomaly detection, with the hyperparameters ‘nu’ and ‘gamma’ optimized to achieve the highest F1-score. Additionally, a baseline Isolation Forest is implemented, where the ‘contamination’ hyperparameter is adjusted to maximize the F1-score as well.

A CNN-LSTM model, adapted from Abdallah et al. (2021), is also included in the comparison. In this model, the CNN layers extract spatial features, which are then passed to LSTM layers to capture temporal patterns and predict the time series. Anomalies are detected by thresholding the prediction errors, with the 95th quantile of the training dataset serving as the threshold.

Additionally, we include a conventional CNN-LSTM autoencoder, structured as shown in Fig. 1, and a conformal CNN-LSTM autoencoder without Temporal Quantile Adjustment (TQA) in the comparison. This allows us to evaluate the impact of conformal prediction and TQA on anomaly detection performance and false alarm rate guarantees. For the conventional CNN-LSTM model, the calibration set was excluded, as conformal prediction was not applied, and the 0.95 quantile of nonconformity scores from the training set was used as the anomaly detection threshold. In the conformal autoencoder model without TQA, the anomaly detection threshold was set to the 0.95 quantile of nonconformity scores from the calibration set as per the conformal prediction pipeline, but without employing TQA to dynamically adjust the threshold. The comparison results are presented in Table 2, with the best performance for each metric highlighted in bold and the second-best performance indicated with underlines.

As shown in Table 2, our proposed conformal autoencoder with TQA attains the highest Precision and F1 scores among the all methods, while maintaining a Recall and AUROC above 0.92. These results demonstrate the model’s robust anomaly detection capabilities, indicating its ability to accurately distinguish between normal and abnormal data. Compared to methods such as PCA (T^2 and SPE), k-NN, CNN-LSTM, conventional autoencoder, and conformal autoencoder without TQA, which rely on the 95% quantile of certain metrics from the training or calibration set for anomaly thresholding, the proposed conformal autoencoder with TQA is the only approach that guarantees a false alarm rate below the desired significance level (0.05). This is because other methods lack rigorous uncertainty treatment and statistical guarantees, particularly under distribution shifts in time series data. While both the conventional autoencoder and the conformal autoencoder without TQA applied a 0.05 significance level using nonconformity scores from the training and calibration sets, they failed to maintain a false alarm rate below 0.05. In contrast, the conformal autoencoder with TQA meets this requirement, ensuring robust anomaly detection and reliable temporal coverage even in the presence of distribution shifts.

Table 2. Evaluation metrics of conformal anomaly detector and benchmark approaches.

Models	Precision	Recall	F1	AUROC	False alarm rate
PCA (T^2 and SPE)	0.7599	0.9225	0.8333	0.8922	0.1382
k-NN	0.8863	**0.9716**	0.9270	0.9563	0.0591
OC-SVM	0.6879	0.9462	0.7967	0.8714	0.2035
Isolation Forest	0.8972	0.9086	0.9028	0.9296	0.0494
CNN-LSTM	0.8826	0.9200	0.9009	0.9600	0.0580
Conventional autoencoder	0.8802	0.9327	0.9057	**0.9635**	0.0602
Conformal autoencoder without TQA	0.8759	0.9327	0.9034	**0.9635**	0.0627
Conformal autoencoder with TQA	**0.9580**	0.9231	**0.9403**	0.9520	**0.0192**

Additionally, the P-value derived from the conformal prediction pipeline provides a quantitative measure for anomaly detection. Since an independent calibration set (separate from the training set) is used to generate P-values, a higher P-value indicates a greater likelihood that a test instance is from the in-distribution (normal) data, making it a valuable indicator for anomaly detection. We illustrate the nonconformity scores and their corresponding P-values for the test instances in our experiment, as shown in Figure 2. Because test instances near the anomaly detection threshold cluster closely around it, distinguishing the degree of anomaly becomes challenging. To address this, we propose a new metric, -log(P-value ratio), which can be calculated as follows.

$$-\log\,(\mathrm{P} - \mathrm{values\ ratio}) =- \,log\left(\frac{P_{test,\,t}}{P_{threshold,\,t}}\right) \quad (18)$$

where $P_{test,\,t}$ and $P_{threshold,\,t}$ are the P-values of the test instance and the anomaly detection threshold at time *t*, respectively. The demonstration of the -log(P-value ratio) can be found at the bottom of Figure 2. In practice, real-time monitoring of P-values and the use of related metrics, such as 1-$P_{test,t}$ and -log(P-value ratio), can provide valuable insights into the anomalous trends in time-series data.

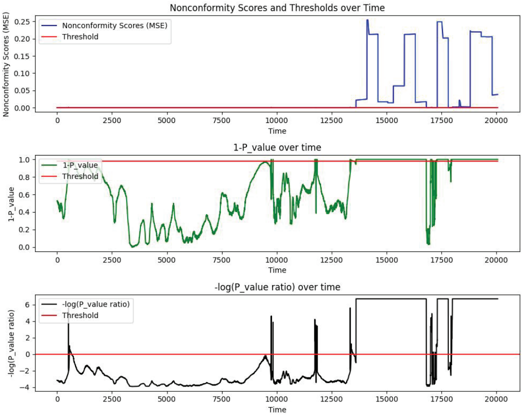

Fig.2 P_value over time generated by the conformal deep autoencoder.

Additionally, we investigate the proposed conformal autoencoder's ability to guarantee controllable false alarm rates by adjusting different pre-defined significance levels within the conformal prediction pipeline. Table 3 presents a comparison of anomaly detection performance across various significance levels, ranging from 0.01 to 0.15, to ensure the desired guarantee of false alarm rate. The results indicate that all experiments meet the required false alarm rate guarantees, with the false alarm rate increasing as the configured significance level rises. Moreover, all experiments achieved F1 scores exceeding 0.87, showcasing the robustness of the conformal autoencoder in delivering strong anomaly detection performance while ensuring controllable false alarm rate guarantees.

Table 3. Comparison between conformal anomaly detection with different significance levels.

Significance level	Precision	Recall	F1	AUROC	False alarm rate
$\alpha = 0.15$	0.8091	0.9414	0.8703	0.9181	0.1053
$\alpha = 0.1$	0.8816	0.9323	0.9062	0.9365	0.0594
$\alpha = 0.05$	0.9580	0.9231	0.9403	0.9520	0.0192
$\alpha = 0.02$	0.9888	0.9152	0.9506	0.9552	0.0049
$\alpha = 0.01$	0.9939	0.9087	0.9494	0.9530	0.0026

5. Conclusions

This study addresses the limitations of unsupervised learning in anomaly detection for industrial control systems, where false positives may lead to unnecessary shutdowns or maintenance. We propose a novel conformal anomaly detection approach that integrates conformal predictions with a CNN-LSTM autoencoder, effectively capturing both spatial and temporal features in time-series data. To further enhance temporal coverage in the presence of distribution shifts, a Temporal Quantile Adjustment (TQA) method is incorporated into the conformal prediction pipeline. Evaluation on a public dataset, along with comparisons to several benchmark methods, demonstrates the proposed model's superior performance in achieving high anomaly detection accuracy while providing robust guarantees for controllable false positive rates.

Acknowledgement

This work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25206422) and the National Natural Science Foundation of China (Grant No. 62406269).

References

Abdallah, Mahmoud, Nhien An Le Khac, Hamed Jahromi, and Anca Delia Jurcut. A hybrid CNN-LSTM based approach for anomaly detection systems in SDNs. In Proceedings of the 16th International Conference on Availability, Reliability and Security, pp. 1-7. 2021.

Alhaidari, Fahd A., and Ezaz Mohammed Al-Dahasi. New approach to determine DDoS attack patterns on SCADA system using machine learning. In *2019 International conference on computer and information sciences (ICCIS),* pp. 1-6. IEEE, 2019.

Angelopoulos, Anastasios N., and Stephen Bates (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511.*

Anton, Simon D. Duque, Sapna Sinha, and Hans Dieter Schotten. Anomaly-based intrusion detection in industrial data with SVM and random forests. In *2019 International conference* *on software, telecommunications and computer networks (SoftCOM)*, pp. 1-6. IEEE, 2019.

Hashim, Hafiz, Paraic Ryan, and Eoghan Clifford (2020). A statistically based fault detection and diagnosis approach for non-residential building water distribution systems. *Advanced Engineering Informatics 46:* 101187.

Kravchik, Moshe, and Asaf Shabtai. Detecting cyber attacks in industrial control systems using convolutional neural networks. In *Proceedings of the 2018 workshop on cyber-physical systems security and privacy*, pp. 72-83. 2018.

Kumar, Vikas, Vishesh Srivastava, Sadia Mahjabin, Arindam Pal, Simon Klüttermann, and Emmanuel Müller. Autoencoder Optimization for Anomaly Detection: A Comparative Study with Shallow Algorithms. In *2024 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8. IEEE, 2024.

Laso, Pedro Merino, David Brosset, and John Puentes (2017). Dataset of anomalies and malicious acts in a cyber-physical subsystem. *Data in brief 14:* 186-191.

Li, Dan, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. *In International conference on artificial neural networks,* pp. 703-716. Cham: Springer International Publishing, 2019.

Lin, Zhen, Shubhendu Trivedi, and Jimeng Sun (2022). Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems 35:* 31017-31030.

Nemani, Venkat, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu (2023). Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing 205:* 110796.

Perales Gómez, Ángel Luis, Lorenzo Fernández Maimó, Alberto Huertas Celdrán, and Félix J. García Clemente (2020). Madics: A methodology for anomaly detection in industrial control systems. *Symmetry 12:* 1583.

Shang, Wenli, Jiawei Qiu, Haotian Shi, Shuang Wang, Lei Ding, and Yanjun Xiao (2024). An Efficient Anomaly Detection Method for Industrial Control Systems: Deep Convolutional Autoencoding Transformer Network. *International Journal of Intelligent Systems, no. 1 (2024):* 5459452.

Scikit,"2.7. Novelty and Outlier Detection." Accessed February 27, 2025. https://scikit-learn.org/stable/modules/outlier_detection.html#.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world.* Vol. 29. New York: Springer, 2005.

Wu, Yulei, Hong-Ning Dai, and Haina Tang (2021). Graph neural networks for anomaly detection in industrial Internet of Things. *IEEE Internet of Things Journal 9:* 9214-9231.

Yuan, Shuaiqi, Ming Yang, and Genserik Reniers (2024). Integrated Process Safety and Process Security Risk Assessment of Industrial Cyber-Physical Systems in Chemical Plants. *Computers in Industry 155* : 104056.

Yang, Tengfei, Yuansong Qiao, and Brian Lee (2024). Towards trustworthy cybersecurity operations using Bayesian Deep Learning to improve uncertainty quantification of anomaly detection. *Computers & Security 144:* 103909.

Zhang, Yuxin, Yiqiang Chen, Jindong Wang, and Zhiwen Pan (2021). Unsupervised deep anomaly detection for multi-sensor time-series signals. *IEEE Transactions on Knowledge and Data Engineering 35:* 2118-2132.

Zhao, Ming, Jingchao Chen, and Yang Li. A Review of Anomaly Detection Techniques Based on Nearest Neighbor. In *Proceedings of the 2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*. https://doi.org/10.2991/cmsa-18.2018.65.