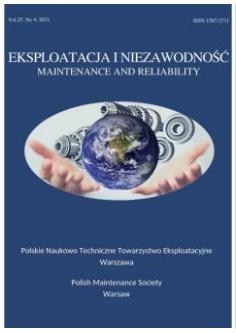


Eksploatacja i Niezawodnosc – Maintenance and Reliability
Volume 25 (2023), Issue 4
journal homepage: http://www.ein.org.pl

Article citation info:
Zhu X, Sun J, Hu H, Li C, A Semi-Supervised Siamese Network for Complex Aircraft System Fault Detection with Limited Labeled Fault Samples, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2023: 25(4) http://doi.org/10.17531/ein/174382

A Semi-Supervised Siamese Network for Complex Aircraft System Fault Detection with Limited Labeled Fault Samples

Indexed by: Web of Science Group

Xinyun Zhu[a], Jianzhong Sun[a,*], Hanchun Hu[a], Chunhua Li[a]

[a] Nanjing University of Aeronautics and Astronautics, China

Highlights

- A novel semi-supervised architecture is proposed high-reliability systems fault detection.
- A comprehensive loss function is employed to achieve the accurate reconstruction of normal samples and the effective separation of fault samples.
- A novel sample pairing strategy is proposed to address the issue of limited labeled fault samples compared to unlabeled data.
- The proposed method is validated with real airline QAR data.

Abstract

Health monitoring and fault detection of complex aircraft systems are paramount for ensuring reliable and efficient operation. The availability of monitoring data from modern aircraft onboard sensors provides a wealth of big data for developing deep learning-based fault detection methods. However, aircraft onboard systems typically have limited labeled fault samples and large amounts of unlabeled data. To better utilize the information contained in limited labeled fault samples, a deep learning-based semi-supervised fault detection method is proposed, which leverages a small number of labeled fault samples to enhance its performance. A novel sample pairing strategy is introduced to improve algorithm performance by iteratively utilizing fault samples. A comprehensive loss function is employed to accurately reconstruct normal samples and effectively separate fault samples. The results of a case study using real data from a commercial aircraft fleet demonstrate the superiority of the proposed method over existing techniques, with improvements of approximately 16.7% in AP, 9.5% in AUC, and 19.2% in F1 score. Ablation studies confirm that performance can be further improved by incorporating additional labeled fault samples during training. Furthermore, the algorithm demonstrates good generalization ability.

Keywords
fault detection, semi-supervised , aircraft system, flight data, time-series data

1. Introduction

In the context of Industry 4.0, the Industrial Internet of Things (IoT) provides an opportunity for the extensive utilization of numerous sensors to continuously monitor aircraft systems, thereby enhancing efficiency, safety, and security[30]. Predictive maintenance approaches aim to minimize Maintenance, Repair, and Overhaul(MRO) downtime and improve safety by identifying component failure precursors in the time-series data obtained from aircraft onboard sensors[8, 10]. Modern commercial aircraft are equipped with an Airplane Condition Monitoring System (ACMS) capable of collecting a wide range of flight data, including operating conditions, status, and performance data during the operation of the aircraft system, which can be utilized for system and component health monitoring and prognostics[52]. Aircraft systems are designed and certified to operate in various challenging environments due to their high safety standard[17, 23]. Aircraft systems rarely fail

(*) Corresponding author.
E-mail addresses: X. Zhu (ORCID: 0000-0002-1813-6710) zhuxinyun@nuaa.edu.cn, J. Sun (ORCID: 0000-0003-1806-7388) sunjianzhong@nuaa.edu.cn, H. Hu acfun0218@164.com, C. Li nuaaccalch@nuaa.edu.cn,

in the whole life cycle leading to aircraft operators accumulating large amounts of unlabeled ACMS data and only a few labeled fault samples. Therefore, effectively utilizing the unbalanced aircraft ACMS and maintenance data for predictive maintenance of aircraft systems is a current research focus[38, 45].

Practical fault detection algorithms are essential for implementing predictive maintenance strategies in real-world applications. Classical fault detection methods, such as linear model-based method [50], distance-based method [1], density-based method[6], and support vector machine[47], have limitations in analyzing multivariate time-series data[9]. Deep learning has proven to be highly effective in reducing the dimensionality of high-dimensional data and learning features in sequential data. Its ability to learn complex data dynamics without assuming underlying patterns makes it a highly attractive choice for time-series fault detection[11, 59].

Fault detection methods based on deep learning can be categorized into three types based on the availability of labeled samples during the training stage: supervised methods[24, 32, 54], unsupervised methods[13, 14, 57], and weakly supervised methods[34, 40]. Obtaining complete, accurate, and exact labels for fault detection tasks can be expensive and challenging due to the high cost and difficulties in data annotation[26]. So in real-world datasets, only a portion of the data will be labeled, as shown in Figure 1.

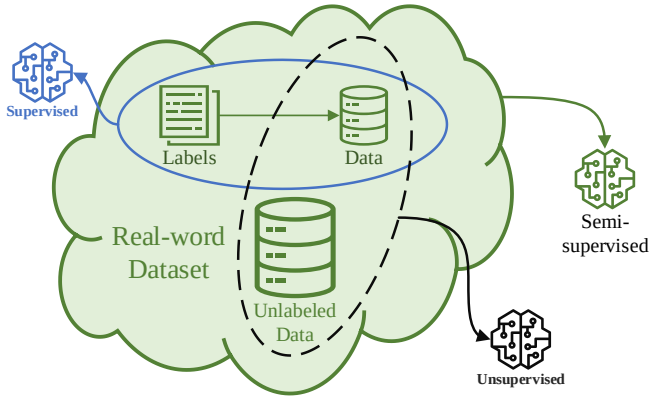


Fig. 1. The usage of data in three types of fault detection methods.

Supervised fault detection methods can utilize labeled data to achieve satisfactory performance usually only when the number of fault data is sufficient and labeled samples are relatively balanced. Unsupervised algorithms miss labels that contain the most crucial information. Therefore, semi-supervised methods as a subset of weakly supervised methods that can leverage limited fault data in extremely unbalanced labeled data scenarios have garnered increasing attention [15, 29].

Deep learning has been widely used in aircraft system fault detection because of its excellent performance in dealing with nonlinear problems and modeling complex nonlinear dynamic systems. Dong integrated deep learning and transfer learning(TL) to propose a TL-enhanced deep learning scheme for aircraft icing and actuator/sensor fault detection and identification using simulation data [12]. Ning et al. propose a Long Short-Term Memory network-based Autoencoder (LSTM-AE) method, which utilizes raw time-series data from heterogeneous sensors for fault detection and classification of complex aircraft systems[38]. Shen et al. proposed an aircraft hydraulic fault diagnosis method based on Empirical Mode Deposition (EMD) and Long Short-Term Memory(LSTM) to eliminate noise interference and adapt to the actual noise environment[48]. Liu et al. based deep learning methods to extract health indicators from raw sensor data of aircraft systems[30]. Zhao et al. used deep neural networks to develop a robust (accurate, scalable, explainable, and interpretable) fault detection scheme for aircraft air data sensors[61]. These methods are supervised[12, 38, 48] and unsupervised learning[30, 61]. The weakly supervised algorithm has been applied in the context of fault detection for airborne systems; however, its application remains relatively less common compared to supervised and unsupervised methods. Li et al.[29] introduced a semi-supervised augmented deep sparse autoencoder (ADSAE) for gear fault detection in aerospace industry.

The goal of this study is to propose a deep learning-based semi-supervised fault detection approach for aircraft systems that is sufficiently robust and reliable in real datasets. Specifically, limited labeled fault samples are utilized during the training phase, resulting in a significant reduction in false positive rates while simultaneously enhancing the accuracy of fault detection, rendering the proposed method appropriate for practical deployment. Differing from typical AE models that solely consider reconstruction errors, our training phase incorporates a comprehensive evaluation of the reconstruction error, the contrastive error, and the partial contrastive error. The

Euclidean distance with necessary adjustments applied is used to quantify contrastive errors between the latent vectors of input pairs. Additionally, the partial contrastive error is employed to impose larger reconstruction errors on anoamly data samples, as a gradient inversion in the training process. Furthermore, a novel anomaly scoring method is introduced. In this scoring scheme, both reconstruction errors and embedding distances are simultaneously employed. Leveraging the advantages of Autoencoder(AE) in data reconstruction, reconstruction errors can mitigate the effect of unexpected fault samples within unlabeled data being classified as normal by default.

The main contributions of this article are the proposed semi-supervised architecture for complex aircraft system fault detection, wherein a novel sample pairing strategy and loss function are proposed to address the issue of imbalanced data in high-reliability aircraft systems with limited labeled fault samples.

(1) A semi-supervised architecture is proposed to address the issue of imbalanced data in high-reliability aircraft systems for fault detection. This approach allows for improved algorithm performance using only a small number of labeled fault samples.

(2) A novel sample pairing strategy is proposed to address the issue of limited labeled fault samples compared to unlabeled data. This strategy involves repeatedly extracting fault samples and pairing them with unlabeled data. The case results demonstrate the significant enhancement of algorithm performance achieved through the proposed novel sample pairing strategy.

(3) A comprehensive loss function is employed to achieve the accurate reconstruction of normal samples and the effective separation of fault samples in the latent space.

This approach allows for improved algorithm performance using only a small number of labeled fualt samples. A case study on a real-world dataset from a commercial aircraft fleet demonstrates the effectiveness and generalization ability of the proposed method. And the case study results show that the proposed architecture outperforms the state-of-art approaches evaluated on the real-world dataset. The rest of this article is organized as follows: Section 2 summarizes the related work of fault detection algorithms based on deep learning. Section 3 introduces the proposed semi-supervised fault detection framework. Section 4 demonstrates the method's effectiveness and generalization ability through a case study on a real dataset. Section 5 discusses the shortcomings of the proposed method, possible solutions, and the follow-up research direction. Finally, a summary of the work is given in Section 6.

2. Related Works

2.1 Deep Supervised Fault Detection

Deep supervised fault detection algorithms have been extensively investigated and demonstrated to outperform traditional strategies. However, these methods typically require a large number of labeled samples, which can be challenging in practical applications. To address this issue, researchers have resorted to using data from test rigs, i.e., simulated data, as a substitute for practical operational data[18]. For instance, Jiang et al. proposed a Stacked Multilevel-Denoising Autoencoder (SMLDAE) based on vibration signals for predicting multiple gearbox faults [25]. Canizo et al. combined Convolutional Neural Networks(CNN) and Recurrent Neural Networks(RNN) to develop a supervised multi-head CNN-RNN anomaly detection model for time-series data derived from a physical model [7].

Additionally, a substantial body of literature has utilized the Commercial Modular Aero-Propulsion System Simulation(CMAPSS) dataset to design numerous engine fault detection and health management methods. Che et al. developed a Prognostic and Health Management(PHM) method for aircraft by combining multiple deep learning algorithms, including supervised Deep Belief Networks(DBN)[10]. Moghadham et al. proposed a neuro-inspired computational model for fault diagnosis through supervised transfer learning[36].

While deep supervised methods have shown satisfactory performance on the test rig and simulation data, the detection accuracy may not be reliable for scenarios not present in the training data, including fault and non-fault scenarios[43]. Furthermore, the imbalanced distribution of labeled samples may also affect the algorithm performance.

2.2 Deep Unsupervised Fault Detection

In recent years, unsupervised fault detection algorithms based on deep learning have gained more attention than supervised algorithms due to their low requirements for labeled data[49]. CNN-based fault detection algorithms have shown promising

results in this field[2, 20, 58]. For instance, Ince et al. proposed a 1-D CNN that enables fast and accurate real-time motor fault detection by analyzing raw data directly [21]. Plakias et al. proposed an Attentive Dense Convolutional Neural Network (ADCNN) that combines dense convolutional blocks and attention mechanisms to detect and identify rolling bearing faults with less training data[42]. Mitra et al. optimized a 1-D CNN using Particle Swarm Optimization (PSO) to achieve accurate transmission line fault detection[35]. Zeiser et al. proposed a high potential online anomaly detection solution based on a combination of Wasserstein Generative Adversarial Networks(WGAN) and encoder CNN[56]. However, CNN treats original time-series data as a spatial distribution of static data, leading to a significant loss of its time-dependent information[30].

In contrast, LSTM is a more mature algorithm for processing time-series data, initially developed in the early 1990s[19]. It has been widely used in fault detection of time-series data and is a variant of the traditional RNN[4] that overcomes the issue of vanishing and exploding gradients. Jalayer et al. proposed a novel Convolutional Long Short Term Memory (CLSTM) for fault detection and diagnosis of rotating machinery[22]. Kłosowski et al. use of the LSTM network to solve the tomographic inverse problem[27]. Belagoune used LSTM to model the spatiotemporal sequences of high-dimensional multivariate features, enabling fault detection of the power system[3]. Zhi et al. combined the CNN-LSTM model to mine the hidden features of processed sensor data to realize fault identification[62]. On the other hand, AE and its variants, as unsupervised learning frameworks, can automatically learn high-level representations directly from complex heterogeneous data[29, 33, 51]. Therefore, researchers have combined the advantages of LSTM and AE algorithms to develop more effective algorithms for fault detection[28, 30, 37, 38].

While unsupervised algorithms can learn detectors from unlabeled data in the absence of prior knowledge, they may not take full advantage of the small number of labeled samples and may suffer from performance degradation in practical application. Hence, semi-supervised algorithms that can make use of a small number of labeled samples may provide a good solution.

In summary, acquiring complete, precise, and accurate labels for supervised fault detection tasks is a difficult and costly task due to the challenges and expenses of data annotation[11], while the unsupervised approaches do not take full advantage of small numbers of labeled samples and may suffer from performance degradation. Therefore, researchers have developed weakly supervised fault detection methods to effectively utilize the limited yet valuable fault samples under incomplete, inexact, and inaccurate supervision[26, 64].

The Siamese network is a deep learning network that uses two or more identical subnets with the same architecture and sharing the same parameters and weights. Koch et al. introduced Siamese Networks to semi-supervised anomaly detection for the first time, which subsequently garnered the attention of other researchers[5, 34, 55]. Castellani et al. attempted to use the Siamese Autoencoder (SAE) to address semi-supervised fault detection[9]. The Siamese network performs well in these tasks because shared weights mean fewer parameters need to be learned during training, and they can produce good results with relatively small amounts of training data. Additionally, the Siamese Networks can be retrained very efficiently as soon as new labeled data are available, making it ideal for refinement during usage as well as for adjusting to drift and reconfigurations of the monitored machinery during operation. Although researchers have developed some weakly supervised fault detection algorithms, most of these methods are designed for relatively simple structures compared to aircraft systems. Moreover, most of the abovementioned methods are still validated through test rigs and simulation data. However, due to the uncertainty of the flight environment and the complexity of aircraft systems, experimental data is difficult to reflect the true working condition of the system.

Hence, this paper introduces a novel semi-supervised fault detection algorithm based on Siamese Networks for aircraft systems, demonstrating its effectiveness through validation on a real-world flight dataset.

3. SSLA for Fault Detection

It is noteworthy that in the task of semi-supervised anomaly detection, the actual annotations of labeled normal samples are not available. Nonetheless, the vast majority of unlabeled samples are normal, and fault samples are exceedingly uncommon in aircraft systems. All unlabeled training samples

are deemed normal to ensure enough labeled normal samples. This strategy has demonstrated satisfactory performance in [39, 63].

3.1 Semi-Supervised Fault Detection Scheme

The proposed fault detection scheme comprises two phases: a training phase and a regular operation phase, as illustrated in Fig. 2.

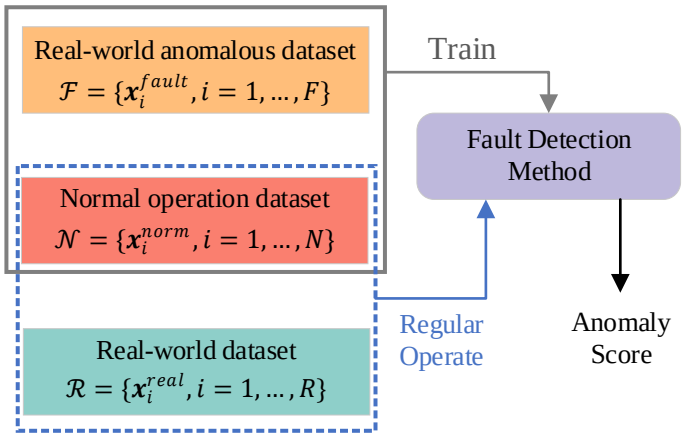


Fig. 2. The overall pipeline of the proposed scheme.

In the training phase, the fault detection model (indicated by the solid gray rectangle) is trained using a combination of two datasets: a normal operation dataset $\mathcal{N}$, which is unlabeled, and a small dataset $\mathcal{F}$ of labeled fault samples. During the regular operation phase (indicated by the dashed blue rectangle), a real-world dataset $\mathcal{R}$ generated from the real word asset is fed to the trained fault detection model along with the normal operation dataset $\mathcal{N}$. An anomaly score(AS) is then calculated for each input data sample, and an example is considered anomalous if its score exceeds a predefined threshold.

3.2 SSLA Network

The Siamese Networks consist of two identical basic networks with shared weights that can efficiently determine whether a pair of input data samples come from the same distribution. This paper proposes a Semi-supervised Siamese LSTM-AE (SSLA) Network, which directly operates on raw time series data, with the encoder and decoder symmetrically chosen.

The proposed structure of the SSLA based on two LSTM-AE basic networks with shared weights is illustrated in Fig. 3. The network consistently assesses a pair of data samples $(\boldsymbol{x}^{norm}, \boldsymbol{x}^{rand})$, whereby the initial sample is always sourced from the normal dataset $\boldsymbol{x}^{norm} \in \mathcal{N}$, and the second sample is randomly selected from either the normal or fault dataset $\boldsymbol{x}^{rand} \in \mathcal{N} \cup \mathcal{F}$.

Usually, the labeled fault samples are much less than the normal samples. To generate more data pairs that comprise fault samples, a new sample pairing strategy is proposed, which can extract data from the fault sample set $\mathcal{F}$ repeatedly, up to a specified number of extractions denoted as n.

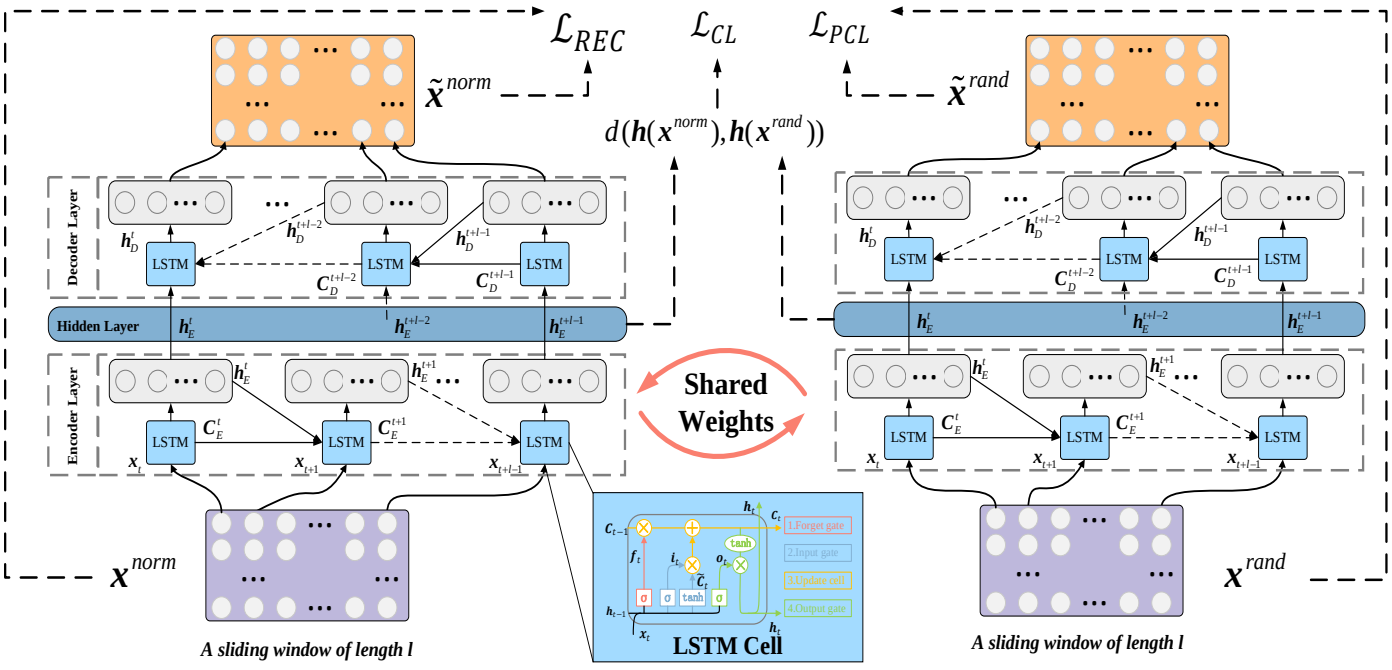


Fig. 3. SSLA structure.

The primary goal of the SSLA is to generate reconstructions of normal data samples with minimal error while simultaneously creating a clear delineation between normal and anomalous data distributions in the latent space. To accomplish this objective, the network must undertake the following actions:

1) Ensure that normal data samples are reconstructed with the utmost precision: $\boldsymbol{x} \simeq \tilde{\boldsymbol{x}}$ for $\boldsymbol{x} \in \mathcal{N}$.

2) Minimize the distance between the latent representations of any two normal data samples: the Euclidean distance $d\big(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x}')\big)$ is small for $\boldsymbol{x} \in \mathcal{N}, \boldsymbol{x}' \in \mathcal{N}$.

3) Generate poor reconstructions of fault data samples, indicated by significantly higher reconstruction error as compared to that of normal data samples: $d(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is large for $\boldsymbol{x} \in \mathcal{F}$.

4) Increase the distance between the latent representations of a normal and a fault data sample: $d\big(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x}')\big)$ large for $\boldsymbol{x} \in \mathcal{N}$ and $\boldsymbol{x}' \in \mathcal{F}$.

Considering the advantages of LSTM in processing time-series data and the learning ability of AE in dimension reduction and feature extraction, the combination of LSTM and AE for multi-sensor time-series anomaly detection is appropriate. The proposed SSLA is built based on two identical LSTM-AE networks. The LSTM-AE model comprises an encoder network, which maps the input data sample $\boldsymbol{x} \in \mathbb{R}$ to a latent representation $\boldsymbol{h}(\boldsymbol{x})$. The decoder network takes the latent representation and reconstructs a data sample in the original space, $\tilde{\boldsymbol{x}} \in \mathbb{R}$. To adapt to the characteristics of the LSTM, a sliding time window is used to segment the flight data. Assuming a sliding time window size of l, the model analyzes a time sequence with a length of l.

In the encoding phase, the encoder accepts a vector of dimension (m, l) as input: $\boldsymbol{x} = (\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_{t+l-1}) \in \mathbb{R}^{m \times l}$. m is the dimension of flight data parameters. The model reconstruction sequence is obtained as $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_{t+1}, \ldots, \tilde{\boldsymbol{x}}_{t+l-1}) \in \mathbb{R}^{m \times l}$.

The goal of the AE model is to reconstruct the input data in an unsupervised manner[53]. The AE consists of two parts: an encoder and a decoder. The encoder maps an input $\boldsymbol{x}$ to a hidden representation $\boldsymbol{h}$ as follows:

$$\boldsymbol{h} = \varphi(\boldsymbol{W} \cdot \boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

where φ is a nonlinear activation function.

The hidden representation $\boldsymbol{h}$ can represent the main features of high-dimensional input data $\boldsymbol{x}$ in low-dimensional space. The decoder transforms the hidden representation $\boldsymbol{h}$ back to the original input as follows:

$$\tilde{\boldsymbol{x}} = \varphi(\boldsymbol{W} \cdot \boldsymbol{h} + \boldsymbol{b}') \tag{2}$$

The parameters $\theta = [\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{W}', \boldsymbol{b}']$ are optimized minimizing an appropriate cost function over the training set, $\boldsymbol{W}, \boldsymbol{W}'$ is weight matrices of networks, $\boldsymbol{b}, \boldsymbol{b}'$ are bias vectors.

The LSTM is a time recursive neural network that can be used to capture the time-series data features in the sequence data learning task. The LSTM cell block diagram is illustrated in Fig. 4.

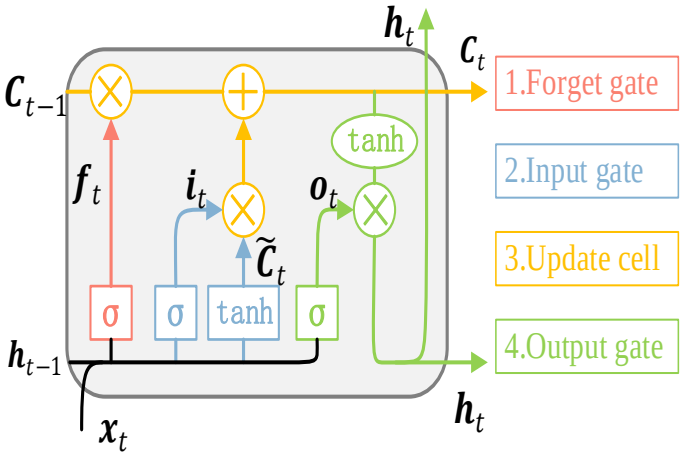


Fig. 4. Scheme of LSTM Cell.

LSTM networks incorporate a gating mechanism that allows the model to decide whether to accumulate or forget certain information regarding the transferred cell state (shown in [16]). The forget gate allows the model to discard useless information from the previous cell state by evaluating the information given by the output of the last step $\boldsymbol{h}_{t-1}$ and the input $\boldsymbol{x}_t$ at the current step t:

$$\boldsymbol{f}_t = \sigma\big(\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f\big) \tag{3}$$

where $\boldsymbol{W}_f$ is the weight matrices, and $\boldsymbol{b}_f$ is the bias vector. The input gate determines which information needs to be updated:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{4}$$

$$\tilde{\boldsymbol{C}}_t = \tanh\,(\boldsymbol{W}_C \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_C) \tag{5}$$

Next, the cell state is updated:

$$\boldsymbol{C}_t = \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t \tag{6}$$

The output of the cell $\boldsymbol{h}_t$ at the current time step is subsequently calculated with the updated cell state $\boldsymbol{C}_t$:

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \tag{8}$$

3.3 Loss Function

The desired outcome can be attained through the implementation of a suitable training regimen for the neural network. A comprehensive depiction of the training regimen is provided in Algorithm 1.

Algorithm1: SSLA Training Procedure

Input: $\mathcal{N}, \mathcal{F}$
Output: $\widetilde{\boldsymbol{x}}^{rand}, \boldsymbol{h}(\boldsymbol{x}^{norm}), \boldsymbol{h}(\boldsymbol{x}^{rand})$
1: Randomly obtain N pairs $(\boldsymbol{x}^{norm}, \boldsymbol{x}^{rand})$ from $\mathcal{N}, \mathcal{F}$
2: Randomly initialize $\boldsymbol{W}_f, \boldsymbol{b}_f, \boldsymbol{W}_i, \boldsymbol{b}_i, \boldsymbol{W}_C, \boldsymbol{b}_C, \boldsymbol{W}_o, \boldsymbol{b}_o$
3: **Repeat**
4: Randomly sample one batch of data pairs
5: Compute the loss $\mathcal{L} = \mathcal{L}_{REC} + \mathcal{L}_{CL} + \mathcal{L}_{PCL}$
6: Perform a gradient descent step update $\left(\boldsymbol{W}_f, \boldsymbol{b}_f, \boldsymbol{W}_i, \boldsymbol{b}_i, \boldsymbol{W}_C, \boldsymbol{b}_C, \boldsymbol{W}_o, \boldsymbol{b}_o\right)$
7: **until** converge
8: **return** $\widetilde{\boldsymbol{x}}^{rand}, \boldsymbol{h}(\boldsymbol{x}^{norm}), \boldsymbol{h}(\boldsymbol{x}^{rand})$

The loss function for training SSLA consists of the following three contributions:

$$\mathcal{L} = \mathcal{L}_{REC} + \mathcal{L}_{CL} + \mathcal{L}_{PCL} \tag{9}$$

where each contribution is calculated as sums over N pairs of data $(\boldsymbol{x}^{norm}, \boldsymbol{x}^{rand})$ and is given by the following expressions, N is the number of $\boldsymbol{x}, \boldsymbol{x} \in \mathcal{N}$.

1) Reconstruction Loss: The mean square error (MSE) between input and reconstruction of operational data, serving as a typical loss function of AE:

$$\mathcal{L}_{REC} = mse(\widetilde{\boldsymbol{x}}^{norm}, \boldsymbol{x}^{norm}) \tag{10}$$

2) Contrastive Loss: The Euclidean distance between the latent vectors of a given input pair, which undergoes a local modification:

$$\mathcal{L}_{CL} = \frac{1}{2}\begin{pmatrix} (1-Y)d\left(\boldsymbol{h}(\boldsymbol{x}^{norm}), \boldsymbol{h}(\boldsymbol{x}^{rand})\right)^2 + \\ Y\left\{\max\left(0, k - d\left(\boldsymbol{h}(\boldsymbol{x}^{norm}), \boldsymbol{h}(\boldsymbol{x}^{rand})\right)\right)\right\}^2 \end{pmatrix} \tag{11}$$

This contribution aims to reduce the difference between inputs belonging to the same class ($Y = 0$), while simultaneously increasing the difference between inputs belonging to different classes($Y = 1$).To prevent individual samples from dominating the loss function, a pre-defined constant, $k > 0$, is introduced, which limits the contribution of samples whose distance is greater than the radius defined by k.

3) Partial Contrastive Loss: Imposing a significant reconstruction error for fault data samples is akin to inducing a gradient inversion during the training:

$$\mathcal{L}_{PCL} = \frac{1}{2} Y max\left(0, k - d(\widetilde{\boldsymbol{x}}^{rand}, \boldsymbol{x}^{rand})\right) \tag{12}$$

Owing to the potential to create a significantly extensive training dataset comprising distinctive pairs extracted from a vast normal dataset and a minuscule fault dataset, this approach can deal with extremely unbalanced datasets,$|\mathcal{N}| \gg |\mathcal{F}|$.

3.4 Anomaly Score

After training, the AS for a new real-world data sample $\boldsymbol{x}^{real} \in \mathcal{R}$ is calculated as follows:

$$AS(\boldsymbol{x}^{real}) = \underbrace{\left(\widetilde{\boldsymbol{x}}^{real} - \boldsymbol{x}^{real}\right)^2}_{Reconstruction\ error} + \underbrace{\frac{1}{N'}\sum_{i}^{N'} \left\|\boldsymbol{h}(\boldsymbol{x}_i^{norm}) - \boldsymbol{h}(\boldsymbol{x}^{real})\right\|_2}_{Embedding\ distance} \tag{13}$$

where $\boldsymbol{x}_i^{norm}$ are from a subset of the normal operation dataset from the training phase with $N' \leq N$elements.

The inclusion of reconstruction errors in $AS(\boldsymbol{x}^{real})$ is to mitigate the effect of unlabeled data being classified as normal by default. If $\boldsymbol{x}_i^{norm}$ is an unexpected anomaly sample, this can render the embedding distance incapable of accurately indicating whether $\boldsymbol{x}^{real}$ is anomaly. By combining the reconstruction error, the advantage of AE in data reconstruction is utilized to enhance anomaly detection. The reconstruction error in AS serves as an indicator of whether sample $\boldsymbol{x}^{real}$ is anomaly. In this article, $N' = 10$. Thus, a single new sample can be paired with multiple normal samples.

The final AS for the entire flight is the sum of the AS for all time sequences sliced from the flight:

$$AS = \frac{1}{M}\sum_{i}^{M} AS\left(\boldsymbol{x}_i^{real}\right) \tag{14}$$

where M is the number of time sequences sliced from a flight. Algorithm 2 describes the regular operational process for calculating AS of one flight.

Algorithm2: Regular operation for flight data

Input: $\mathcal{N}$, all the data on an entire flight
Output: AS
1: Segment the flight data using a sliding time window
2: **for** $i = 1\ to\ M$ **do**
3: Randomly obtain N' normal samples from $\mathcal{N}$
4: **for** $j = 1\ to\ N'$ **do**
5: evaluates a pair of data samples $\left(\boldsymbol{x}_j^{norm}, \boldsymbol{x}_i^{real}\right)$
6: **end for**
7: Compute $AS\left(\boldsymbol{x}_i^{real}\right)$
8: **end for**
9: **return** AS

4. Case Study on Reliable Aircraft System

4.1 Dataset Description

4.1.1 Sensor Data of Aircraft Air Conditioning System

The case study is carried out on a real dataset from the Air Conditioning System(ACS) of a single-aisle twin-engine commercial aircraft. Serving as an extensive thermal control system, the ACS effectively regulates various parameters such as temperature, pressure, and humidity, among others, to provide a comfortable working and living environment for both crew and passengers. Extensive on-board sensor monitoring data is acquired and recorded as ACMS data, which can be further analyzed for system condition monitoring in real-time or after the flight. Fig. 5 displays the principal configuration of the studied aircraft ACS.

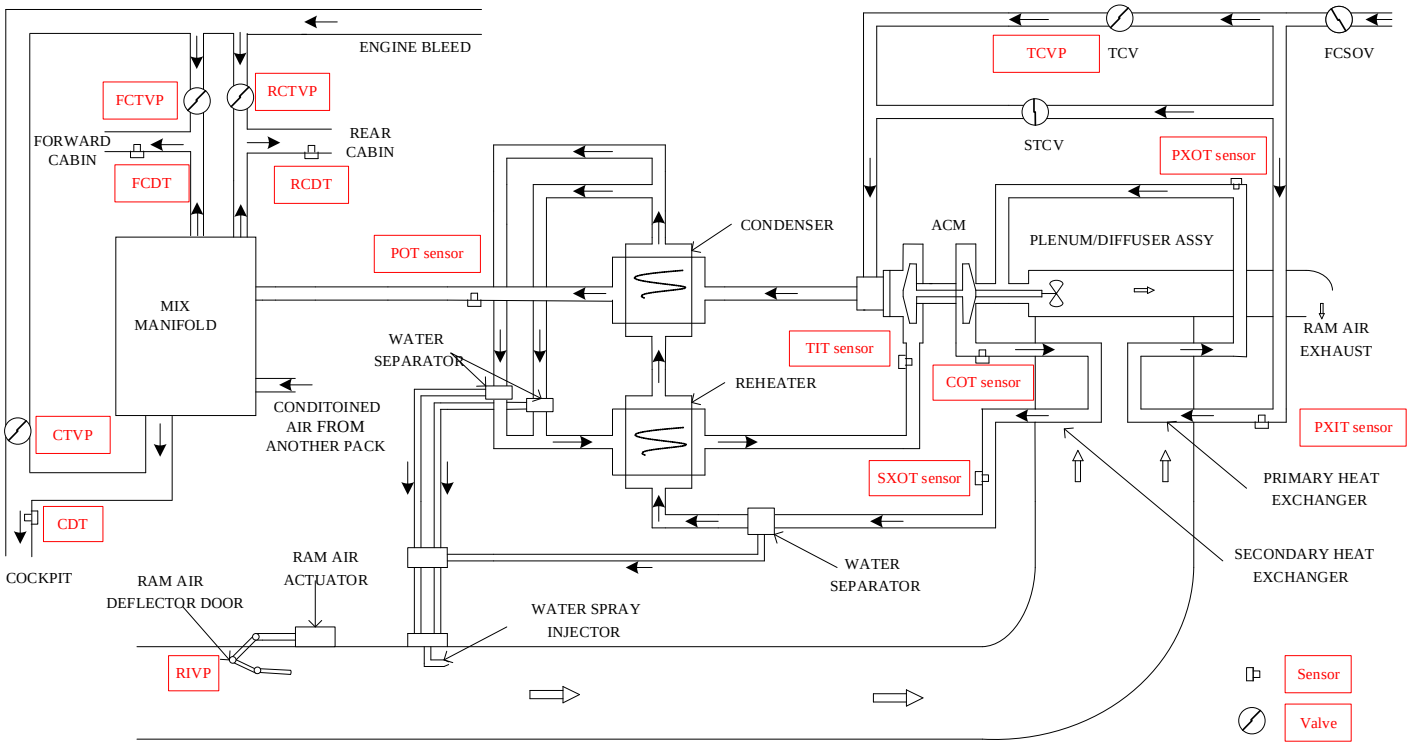


Fig. 5. The configuration of the ACS.

Table 1. ACS parameters recorded in ACMS.

Parameter	unit
Computed Airspeed (Cas)	Ma
Total Air Temperature (Tat)	℃
Cockpit Duct Temp(CDT)	℃
Front Cabin Duct Temperature(FCDT)	℃
Rear Cabin Duct Temperature(RCDT)	℃
Cockpit Trim Valve Position(CTVP)	%
Front Cabin Trim Valve Position(FCTVP)	%
Rear Cabin Trim Valve Position(RCTVP)	%
Pack Outlet Temperature(POT)	℃
Primary Heat Exchanger Inlet Temperature(PXIT)	℃
Primary Heat Exchanger Outlet Temperature(PXOT)	℃
Secondary Heat Exchanger Outlet Temperature(SXOT)	℃
Compressor Outlet Temperature(COT)	℃
Ram Intake Valve Position(RIVP)	%
Temperature Control Valve Position(TCVP)	%
Turbine Inlet Temperature(TIT)	℃

The main parameters utilized in this study are given in Table 1. The ACMS collects these parameters and records them at a frequency of 1Hz for each flight. Fig. 6 displays the PXOT and POT parameters of the ACS during two typical flights of the same aircraft. It can be observed that the PXOT and POT parameters exhibit noticeable fluctuations during the whole flight, which can be mainly attributable to the ACS switching between different functional modes due to changes in flight phases. The complex system consists of a large number of components that closely interact with each other leading to complex dependencies between sensor readings. Furthermore, the ACS consists of many feedback, control, and safety mechanisms, therefore, the simple analysis of the ACS raw sensor parameters may not be sufficient for an effective health monitoring solution, since the redundancy and control mechanisms are able to compensate for a failure. As a consequence, a simple temperature or pressure exceedance-

based health monitoring method may not be sufficient. There is a pressing need to develop an effective fault detection method.

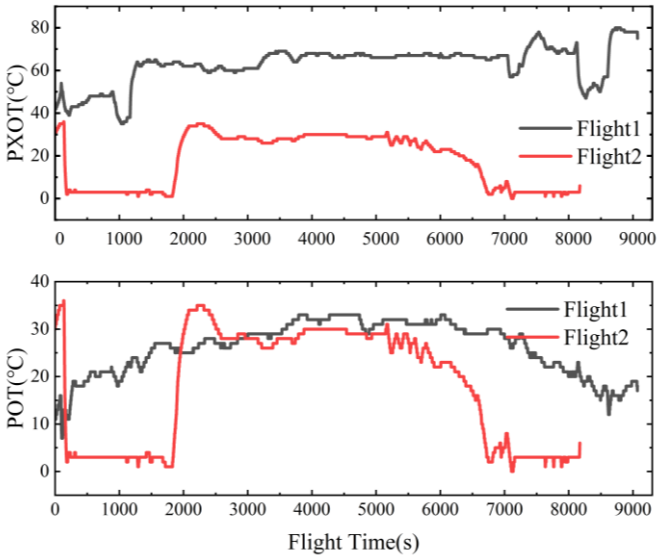


Fig. 6. PXOT and POT of ACS during two flights.

4.1.2 Fault Identification Based on Expert Experience

To begin with, it is necessary to identify the fault data and part of the performance degradation data from the dataset for model training. This procedure primarily relies on the expertise of these professionals and is not governed by specific, standardized criteria or benchmarks. Firstly, maintenance experts initiate their evaluation by checking maintenance logs to identify functional faults or issues reported during previous flights. Subsequently, they focus on key system performance parameters, such as PXOT and POT, among others. These parameters are monitored for any significant deviations from their expected values. And experts consider the overall operational conditions during the flight, including the aircraft's mode of operation, environmental factors, and any relevant external conditions that may have influenced the system's performance. The critical aspect of this process is expert judgment. There isn't a predefined set of numerical thresholds or benchmarks for declaring a performance degradation flight. In conclusion, the identification and categorization of fault and performance degradation flights are inherently subjective and are reliant on the discernment and expertise of maintenance professionals. While this approach lacks specific standards, it ensures a comprehensive evaluation of potential performance issues.

4.1.3 Dataset Preprocessing

The dataset is collected from a fleet comprising four aircraft, encompassing a total of 1049 flights. A total of 4 flights with function faults have been labeled according to the maintenance records. A functional fault refers to a failure mode where a device is unable to continue performing its intended function. Such faults can result in operation interruption with significant economic losses and even pose safety hazards in the context of aircraft. Hence, it is imperative for aircraft operators to detect and repair such faults as early as possible.

After repeated confirmation and discussion with domain experts, a total of 90 performance degradation flights (anomaly flights) prior to functional faults have been manually identified to assess the proposed SSLA method in the following part of the paper. The datasets from the four aircraft are summarized in Table 2.

Table 2. Summary of the dataset.

Aircraft ID	Unlabeled Flights	Performance Degradation	Function Fault
A	253	37	1
B	308	43	1
C	130	8	1
D	264	2	1
Total	955	90	4

The sensor data are from different domains and scales, therefore, they need to be standardized with the z-score method : $Z = (X - \mu)/\sigma$, where μ is the mean of X and σ its standard deviation. In this study, a time window of 30 seconds is adopted, and each input pair comprise two sequences with 16 dimensions and 30 seconds in length. Time sequences are obtained from the flight data using a time window of length l and a 1-second stride. Therefore, the value of M varies for each flight, changing in accordance with the duration of the flight. In the data set used in this paper, M is generally 5000~10000.

4.2 Performance Metrics

Multiple performance metrics are employed to assess the effectiveness of the discussed methods in this study. The result of classification is typically presented in a confusion matrix, as shown in Table 3. Accuracy is the ratio of the total number of correct predictions to the total number of predictions: $Accuracy = (TP + TN)/(TP + FP + TN + FN)$. Precision is the ratio of the total number of positives correctly

predicted to the total number of positives predicted:$Precision = TP/(TP + FP)$. The recall is the ratio of the total number of positives correctly predicted to the total number of positives:$Recall = TP/(TP + FN)$.

Table 3. Confusion metrics.

		Predicted Label	
		Normal	Anomaly
True Label	Normal	True Negative(TN)	False Positive(FP)
	Anomaly	False Negative(FN)	True Positive(TP)

The first metric is the F1 score, which is the harmonic mean of precision rate and recall rate:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{15}$$

The second one is AUC, the area under the receiver operating characteristic curve(ROC). The horizontal axis of the ROC curve represents the False Positive Rate (FPR):$FPR = FP/(FP + TN)$, while the vertical axis represents the True Positive Rate (TPR), that is, the recall rate.

And the third one is average precision (AP), which quantifies the area under the precision-recall curve (PRC). The horizontal axis of the PRC curve represents the recall rate, while the vertical axis represents the precision rate. In the case of highly unbalanced class sizes, it has been demonstrated that PRC is more informative than ROC[46], as it better captures the accurate prediction of the minority class.

To assign a specific label to a given data sample, a threshold (θ_{AS}) is applied to the calculated AS to identify anomalous samples. If the AS of the test data exceeds the threshold, it is considered anomalous. For the sake of clarity, the explicit value of θ_{AS} is required for calculating the F1 score and the confusion matrices. All possible threshold values are considered for other performance measures employed in this study, such as AP and AUC.

4.3 Experiments Result

4.3.1 Experiment Design

This section reports the performance of the proposed semi-supervised approaches compared to other state-of-the-art algorithms. Isolation Forest (IF) is a commonly used technique for anomaly detection that builds an ensemble of isolation trees to identify anomalies as those data points with short average path lengths in the trees[31]. The K-Nearest Neighbor(KNN) method calculates outliers metrics based on the distance to the nearest neighbor[41]. The paper also compares the results of the unsupervised Local Outlier Factor (LOF), which measures data point anomalies based on the local deviation of their densities relative to their neighbors[6]. The unsupervised dimensionality reduction method Principal Component Analysis(PCA)[50]and the LSTM-AE[30] are used as the fault detection technology, using the reconstruction error of each sample as the AS. Deep learning-based supervised algorithms SAE[9] and Multilayer Perceptron(MLP)[44] are also included in the comparison.

The network structure and hyperparameter settings of LSTM-AE are consistent with those in [30]. The MLP includes two hidden layers, each of which comprises 64 neurons. The SAE model is equipped with identical hyperparameters to the SSLA model. The SSLA model consists of two identical LSTM-AE networks. The detailed network layer structure is shown in Table 4, where the architecture and hyperparameters of the LSTM-AE are presented as follows: an input layer, two encoding layers, an embedding layer, two decoding layers, and an output layer. The input and output layers contain a number of neurons equal to the dimensionality of the multivariate data, i.e., 16. The encoding layers consist of 128 and 64 neurons, respectively, the embedding layer has 15 neurons, and the decoding layers consist of 64 and 128 neurons, respectively. The maximum number of iterations is set to 100, and the Adam optimizer is chosen as the optimization algorithm. The remaining hyperparameters are maintained at their default values. Each input sample corresponds to a 30-second time-series data comprising 16 dimensions.

Table 4. Network Layer Structure of LSTM-AE.

Layer	Output shape	Description
Input	(30, 16)	Input layer
LSTM_1	(30, 128)	LSTM encoding layer1
LSTM_2	(30, 64)	LSTM encoding layer2
LSTM_3	(30, 15)	LSTM embedding layer
LSTM_4	(30, 64)	LSTM decoding layer1
LSTM_5	(30, 128)	LSTM decoding layer2
LSTM_6	(30, 16)	LSTM output layer

In accordance with the strategy outlined in Section 3, unlabeled data should be treated as normal data. 80% of the 955 unlabeled data, i.e., 764, along with four function fault flights, are employed to train semi-supervised(SAE, SSLA) and supervised algorithms(MLP). Specifically, for SSLA, the labeled fault sample set $\mathcal{F}$ employed in training consists of 4

fault flights, and the normal sample set $\mathcal{N}$ comprises 764 flights. As for unsupervised algorithms, the set $\mathcal{N}$ is utilized for training. The remaining 20% of unlabeled flights and 90 performance degradation flights are grouped into the set $\mathcal{R}$ for the evaluation of all algorithms. The value that makes F1 the largest is selected as the threshold θ_{AS} within the value range of AS in the set $\mathcal{R}$.

For both the supervised and semi-supervised methods, fault samples are repetitively extracted a total of four times in training phase, denoted by $n = 4$.

The current study utilizes state-of-the-art anomaly detection algorithms sourced from the open-source library PyOD for comparison purposes[60]. The neural networks employed in this analysis are implemented utilizing the Keras library, with the TensorFlow 2.6 back-end. The computational resources leveraged in the training process consist of an Intel Core E5-2640 processor clocked at 2.4 GHz, alongside an NVIDIA Quadro P2200 GPU.

4.3.2 Comparison Analysis

To determine the value of N', exploratory analysis is conducted. N' is varied from 1 to 20. The average detection time for the entire flight data increases proportionally with N'. When $N' = 10$, the average detection time is approximately 35 seconds, but when $N' = 20$, the average detection time increases to around 66 seconds. In the context of a flight lasting several hours, 35 seconds can be considered negligible. The average detection time indicates the real-time detection capability of the SSLA. Following the acquisition of flight data, it can promptly identify faults.

And the performance metrics AP is computed, as illustrated in the Fig. 7.

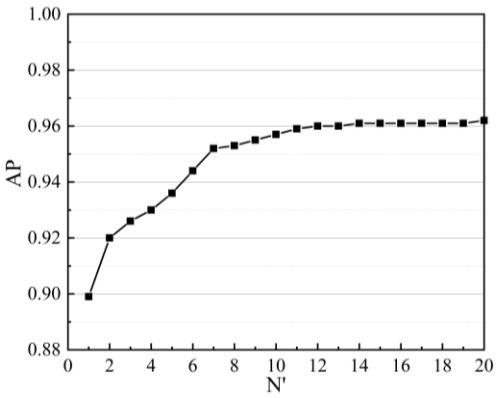


Fig. 7. The impact of N' on AP.

It can be observed that when N' is 10, it approaches its maximum value of 0.957. While further increasing N' may yield marginal improvements in performance, considering the constraints on computational efficiency, $N' = 10$ is suitable.

Fig. 8 depicts the normal and fault instance data, along with their corresponding reconstructions obtained using the SSLA method. Specifically, the top row shows the PXOT, and the bottom row shows the POT. The results demonstrate the effectiveness of the developed method in accomplishing the predetermined objective as outlined in Section 3.1. More precisely, the SSLA method achieves accurate reconstruction of the normal data, whereas the fault instance reconstruction exhibits substantial dissimilarities from the original data. Notably, the data has been standardized prior to the reconstruction process.

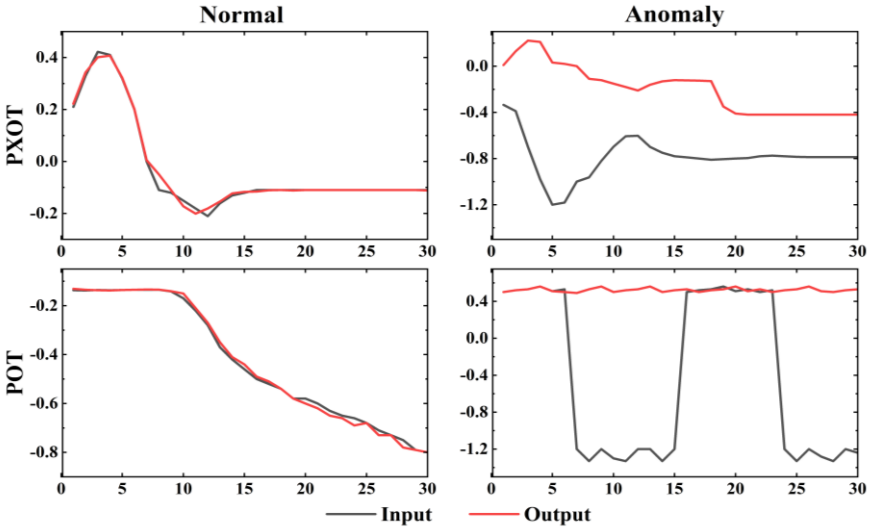


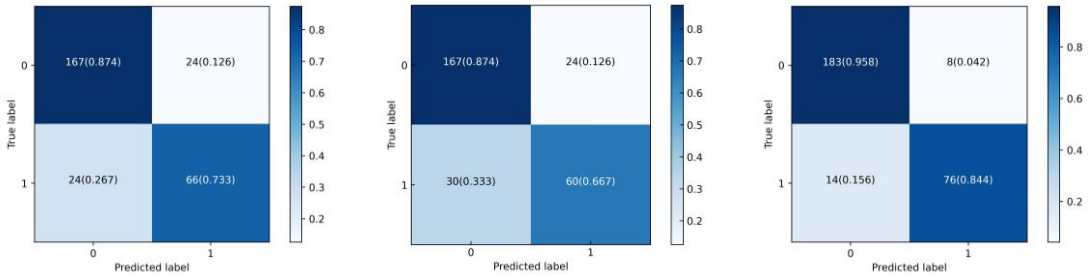
Fig. 8. The instance data and reconstruction.

The final test and evaluation involved 191 normal flights and 90 performance degradation flights from the entire fleet, as illustrated in Section 4.1.2, with the corresponding performance values presented in Table 5. Among the unsupervised methods, the deep-learning-based LSTM-AE exhibited the best performance, with a slightly higher AP score of 0.831, outperforming even the supervised MLP and SAE. This outcome aligns with the anticipated findings, as previous research[1] has reported the algorithm's superior performance for the system under investigation.

The proposed semi-supervised algorithms demonstrate superior performance across all metrics in comparison to unsupervised methods. This outcome is expected due to the utilization of additional valuable information pertaining to actual faults during the training process. The findings of our investigation indicate that the SSLA algorithm yields the best performance among all the algorithms considered, achieving an AP score of 0.957, AUC ROC of 0.975, and F1 score of 0.874. Notably, when trained with the same number of labeled samples, the SSLA algorithm surpasses the unsupervised LSTM-AE by 16.7%, the supervised MLP by 28.8%, and the SAE by 19.8% in the AP score.

Table 5. Performance values of all algorithms.

Train Approach	Algorithm	AP	AUC	F1	Accuracy	Recall	Precision
Unsupervised	PCA	0.614	0.686	0.634	0.762	0.644	0.624
	LOF	0.617	0.742	0.625	0.765	0.611	0.640
	KNN	0.709	0.759	0.681	0.790	0.700	0.663
	iForest	0.555	0.684	0.542	0.712	0.533	0.552
	MCD	0.689	0.704	0.630	0.758	0.644	0.617
	LSTM-AE	***0. 831***	***0.882***	***0.733***	***0.829***	***0.733***	***0.733***
Supervised/ Semi-supervised	MLP	0.829	0.881	0.759	0.840	0.789	0.732
	SAE	0.799	0.812	0.690	0.808	0.667	0.714
	SSLA	***0.957***	***0.975***	***0.874***	***0.922***	***0.844***	***0.904***

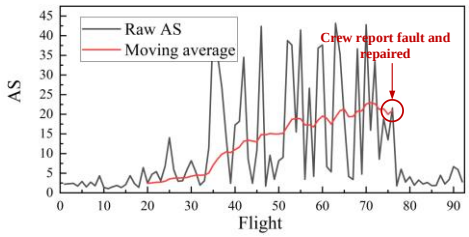


(a) LSTM-AE (b) SAE (c) SSLA

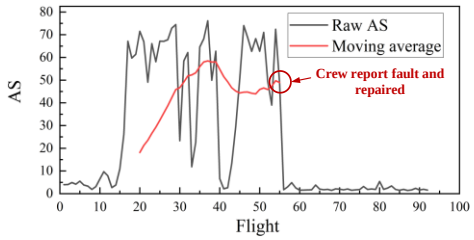
Fig. 9. Confusion matrices of the proposed algorithms for fleet fault detection.

Fig. 9 shows the confusion matrices of the best performing unsupervised algorithm LSTM-AE [see Fig. 9(a)] and semi-supervised SAE [see Fig. 9(b)], SSLA [see Fig. 9(c)]. The numbers in Fig. 9 correspond to the absolute detections, while those given in parentheses represent the in-class percentages. The false positive rate (FPR) decreases from 12.6% to 4.2%, and the false negative rate (FNR) is approximately halved from 26.7% to 15.6% with the shift from unsupervised to semi-supervised training. The SSLA, as the best-performing method, has about 4.2% of FPR and 15.6% of FNR, which is still rather large but already at the edge of being feasible for a real-world application. It is worth emphasizing that during semi-supervised training, only four labeled anomalous samples were repeatedly sampled four times. It can be reasonably anticipated that increasing the number of labeled samples would further reduce the false positive rate (FPR) and false negative rate (FNR).

Fig. 10 shows the variation trend of AS when the SSLA algorithm conducts fault detection on aircraft B[see Fig. 10(a)] and C[see Fig. 10(b)]. The data analyzed is collected from approximately 90 flights around the time of the fault, none of which are utilized in the training process. The raw AS is subjected to a smoothing procedure using a moving average of length 20. Prior to the onset of the fault, AS gradually increases, whereas following the report of the fault and subsequent maintenance, AS returns to pre-fault levels. These findings indicate that AS represents a dependable indicator of ACS health. The SSLA method is anticipated to enable the prognostication of aircraft system faults in the future.



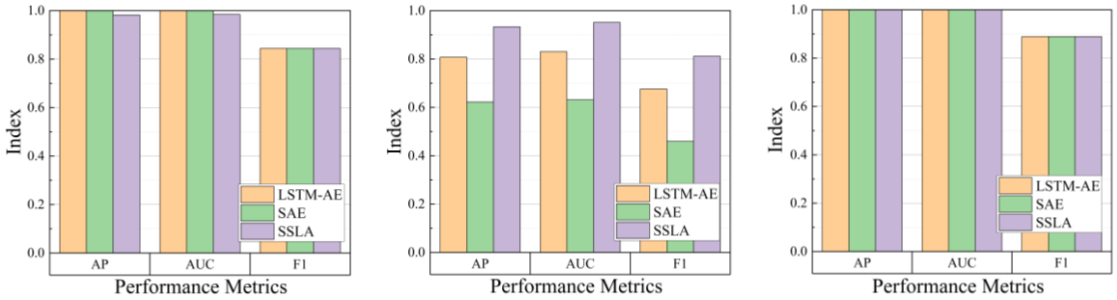
(a) aircraft B



(b) aircraft C

Fig. 10. Computed AS for aircraft B and C.

The entire fleet consists of four aircraft. Fig. 11 shows the fault detection performance of LSTM-AE[see Fig. 11(a)], SAE[see Fig. 11(b)] and SSLA[see Fig. 11(c)] for three aircraft, respectively, while the performance degradation flights of aircraft D are insufficient to be included in this comparison. The three algorithms exhibit optimal detection performance for aircraft A/C. Conversely, the results of LSTM-AE and SAE for aircraft B are terrible, with an AP of not higher than 0.8. However, the proposed SSLA algorithm maintained an AP score of over 0.93, significantly surpassing the other two algorithms. These findings suggest that the SSLA algorithm is robust and capable of delivering excellent detection performance for various faults.



(a) aircraft A (b) aircraft B (c) aircraft C

Fig. 11. Comparison of single aircraft fault detection performance.

4.3.3 Ablation Study

In comparing algorithm performance, the SSLA has demonstrated superior performance compared to the traditional SAE algorithm when trained with a limited number of fault samples. Consequently, an investigation into the impact of the number of input pairs of fault samples on algorithm performance is of interest. The flights with degraded performance are added to the training set $\mathcal{F}$as faulty flights due to insufficient faulty flights.

To this end, sets of differently sized fault samples {1,2,3,4,8,12}, corresponding to 0.13%, 0.27%, 0.38%, 0.52%, 1.05%, and 1.57% of normal training samples, are used to train supervised and semi-supervised algorithms. All fault samples are repeatedly extracted four times, i.e., $n = 4$. As expected, the performance of the three supervised methods generally increases with an increased number of labeled fault samples, as observed in Fig. 12. Notably, none of the three algorithms can achieve superior performance when only one fault sample is used for training, with average precision (AP) hovering around 0.6. In contrast, the proposed SSLA algorithm yields a remarkably high AP score of 0.988 with only eight labeled fault samples.

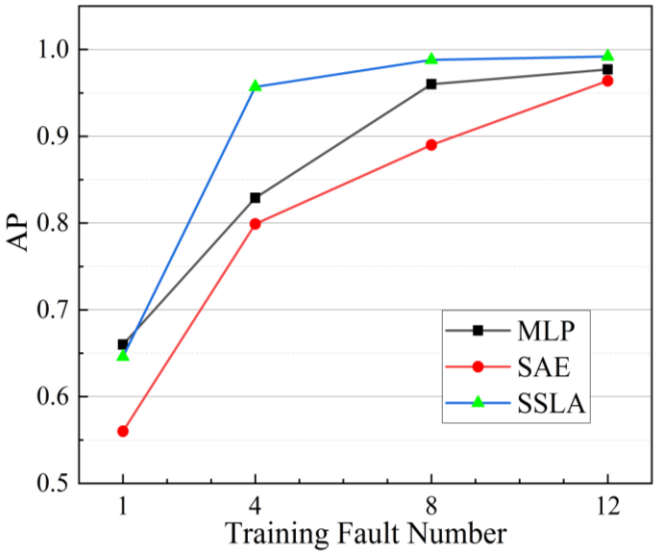


Fig. 12 .Algorithm performance trained with different numbers of fault samples.

The dependency between the number of fault sample repeats ,i.e., n and the performance of SSLA is presented in Fig.

13. Fault samples are extracted at varying repetitions, specifically for 1, 2, 4, and 8 times, denoted as $n = 1,2,4,8$. The algorithm performance is very poor when trained with only one fault sample. However, better performance can be achieved through repeated extraction when the fault set size is 2, which proves the effectiveness of the proposed strategy for repeating fault samples. Notably, as the fault set size increases, such as when using 8 or 12 fault samples, repeated extraction no longer improves the algorithm's performance. Therefore, it is crucial to select an appropriate fault sample size and the number of repeated extractions to optimize the algorithm's performance.

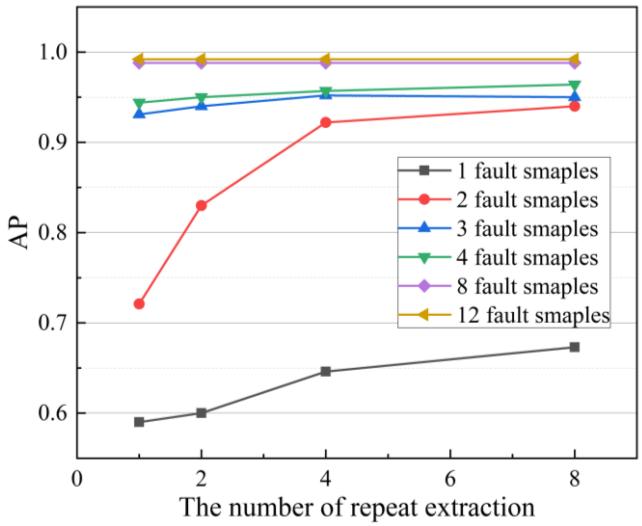


Fig. 13. Algorithm performance trained with different numbers of repeat extraction.

4.3.4 Generalization Ability Analysis in Fleet Data

Aircraft operators require fault detection methods that exhibit strong generalization capabilities to accommodate the diversity of fleet data. To assess the generalization ability of SSLA, an additional validation dataset has been incorporated. This dataset comprises 735 flight data from the ACS of two aircrafts, identified as E and F. The data from these two aircrafts were not utilized during the model training process.

The model was trained using flight data from aircraft A, B, C, and D, with four fault samples repeatedly extracted four times ($n = 4$). The details of the validation dataset are provided in Table 6. The performance of the SSLA method on untrained fault samples from aircraft D and E is presented in Table 7. Remarkably, the method exhibits excellent generalization ability, achieving a total AP of 0.913, thus effectively detecting anomalies in fleet data.

Table 6. Summary of the validation dataset.

Aircraft ID	Unlabeled Flights	Performance Degradation	Function Fault
E	322	17	1
F	413	3	1
Total	735	20	2

Table 7. Performance values of SSLA on aircraft D/E.

Aircraft ID	AP	AUC	F1	Accuracy	recall	precision
D	0.903	0.956	0.941	0.994	0.887	1
E	1	1	1	1	1	1
Total	0.913	0.933	0.952	0.997	0.909	1

5. Discussion

5.1 Evaluation and Limitation of the Method

The most unique feature of this work is the proposal of a novel deep learning-based semi-supervised fault detection algorithm(SSLA). SSLA leverages limited fault data samples during the training process to accurately reconstruct normal data while separating normal samples from fault samples in the latent representation. The algorithm's performance is verified on a real flight dataset in three performance metrics, and it exhibits good robustness to different faults. Moreover, the ablation analysis revealed that increasing the number of fault samples used in training enhances the algorithm's accuracy. These findings imply that SSLA has potential utility in detecting faults in intricate real-world systems.

Previous fault detection methods for complex aircraft systems have mainly been unsupervised methods, with LSTM-AE being a typical example. In comparison, the method proposed in this study can fully utilize the small but high-value fault data accumulated by operators. Moreover, it can provide higher accuracy and robustness, and is more practical.

In our approach, while SSLA contributes to improving the accuracy of fault detection, it is essential to emphasize that the algorithm serves as a valuable tool to assist maintenance experts rather than replacing their expertise. The role of our algorithm is to provide preliminary assessments and alerts based on data-driven patterns and anomalies. Subsequently, human experts, who possess domain knowledge and experience, are responsible for making the final judgment regarding the necessity of maintenance actions. To manage potential false alarms, the two-fold strategy is recommended. Firstly, the algorithm should be continuously refined and update using feedback from

maintenance experts and further training on real-world data. This iterative process aims to reduce false alarms over time as the algorithm becomes more attuned to the specific operational context. Secondly, a tiered alert system can be implemented based on the size of the AS. Maintenance experts would prioritize their attention on alerts with higher severity levels, while lower-severity alerts would undergo additional scrutiny or be flagged for further monitoring, reducing the chances of unnecessary interventions. For missed faults, our approach promotes a proactive maintenance strategy. Flight operators are encouraged to closely monitor the algorithm's performance and continuously validate it against their historical maintenance records. Any missed faults are treated as opportunities for algorithm enhancement and refinement.

In this article, it is presumed in this article that all unlabeled data is normal data, which may not be accurate in real-world scenarios. Complex systems, including feedback, control, and security mechanisms, often mask early fault detection due to redundancy. While some unlabeled data may deviate from the system's normal operation (e.g., performance degradation from coupled failures), such instances are rare in highly reliable aircraft. Stringent maintenance programs ensure system functionality, and any failures prompt immediate maintenance and logging. Consequently, a minimal amount of fault data exists in unlabeled data. To mitigate the impact of anomalies, one approach involves reevaluating unlabeled data using the initial model, removing anomalies, and iteratively retraining the model for improved accuracy.Due to space limitations, this article only discusses the application of SSLA in fault detection. In the forthcoming period, to facilitate the application and implementation of deep learning algorithms in practical engineering predicaments, it is imperative to probe the efficacy of SSLA in the context of prognostics and health management (PHM), specifically, in predicting the remaining useful life (RUL). Predicting RUL enables operators to foresee faults in advance and carry out maintenance activities as promptly as feasible, ensuring aviation safety and curtailing operating expenses.

5.2 Limitations of Deep Learning within the Safety-critical Domain

The potential limitations of deep learning in terms of robustness and interpretability within the safety-critical domain is indeed crucial to be discussing. These discussions are helpful to the practical application of the proposed method in the safety-critical domain and to clarify the research direction. Here, these limitations are delved into and possible pathways for future solutions are provided:

(1)Robustness Challenges:

Deep learning models are vulnerable to adversarial attacks, where small perturbations in input data can lead to misclassification. To address this, ongoing research focuses on adversarial training and robust model architectures. And deep models may struggle to generalize to untrained scenarios. Addressing this challenge involves collecting diverse and representative data, augmenting training sets, and exploring transfer learning techniques.

(2)Interpretability Challenges:

Deep learning models often lack transparency, making it difficult to understand why they make specific predictions. Interpretable models and post-hoc explanation techniques like Local Interpretable Model-Agnostic Explanations(LIME) and SHapley Additive exPlanations(SHAP) are actively researched areas. Enhancing interpretability also involves estimating model uncertainty. Bayesian deep learning and probabilistic models aim to provide more reliable uncertainty estimates.

To address these limitations, interdisciplinary collaboration between machine learning experts, domain specialists, and regulatory bodies is crucial. Moreover, research efforts should continue to focus on developing robust, interpretable deep learning models specifically tailored for safety-critical applications.

6. Conclusion

Health monitoring and fault detection of complex engineering systems, such as aircraft systems, are of paramount importance in ensuring their reliable and efficient operation. However, the use of supervised methods in complex aircraft systems may be constrained by the significant imbalance in labeled samples resulting from insufficient fault data. And unsupervised algorithms may not fully leverage these limited labeled samples and may suffer from degraded performance in practical deployment.

To address the challenge of utilizing limited fault

information for robust fault detection in complex aircraft systems, this paper proposes a Semi-supervised Siamese LSTM-AE (SSLA) framework tailored to multivariate time-series sensor data and can improve detection performance with limited labeled fault samples. The algorithm evaluates a pair of samples to determine whether they are from the same distribution. To address the issue of having significantly fewer labeled fault samples compared to unlabeled data, a novel pairing strategy is proposed that involves repeatedly extracting fault samples and pairing them with unlabeled data. In the model training phase, reconstruction error, contrastive loss, and partial contrastive loss are integrated as the error function. This approach minimizes the reconstruction error of normal data while achieving a significant separation between normal and fault data in the latent space.

The algorithm is comprehensively evaluated using various performance metrics on a real-world dataset from the Air Conditioning System of commercial aircraft. Anomaly scores are constructed using a combination of reconstruction error and embedding distance. Compared to the state-of-the-art fault detection algorithms, superior results are achieved in terms of AP, AUC, and F1 metrics. Specifically, compared to the traditional LSTM-AE approach, the proposed method demonstrated an improvement of 16.7% in AP, 10.5% in AUC, and 19.2% in F1, respectively. Furthermore, the proposed method, SSLA, exhibited robustness and excellent detection capabilities for different faults. The algorithm's performance also improved with an increase in the number of training fault samples. By applying a pre-trained model to an additional validation dataset comprising flight data from two aircraft for fault detection, the proposed method achieved an overall AP of 0.913. This result demonstrates the strong generalization ability of the approach. These findings suggest that the SSLA algorithm can be utilized for fault detection in real-world aircraft systems with few fault samples.

Acknowledgments

This work was supported by the NSFC & CAAC Joint Research Fund (No. U2233204) and Fund of Shanghai Engineering Research Center of Civil Aircraft Health Monitoring (GCZX-2022-02).

Reference

1. Angiulli F, Pizzuti C. Fast Outlier Detection in High Dimensional Spaces. In Elomaa T, Mannila H, Toivonen H (eds): Principles of Data Mining and Knowledge Discovery, Berlin, Heidelberg, Springer Berlin Heidelberg: 2002; 2431: 15–27, https://doi.org/10.1007/3-540-45681-3_2.
2. Basumallik S, Ma R, Eftekharnejad S. Packet-data anomaly detection in PMU-based state estimator using convolutional neural network. International Journal of Electrical Power & Energy Systems 2019; 107: 690–702, https://doi.org/10.1016/j.ijepes.2018.11.013.
3. Belagoune S, Bali N, Bakdi A et al. Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. Measurement 2021; 177: 109330, https://doi.org/10.1016/j.measurement.2021.109330.
4. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 1994; 5(2): 157–166, https://doi.org/10.1109/72.279181.
5. Berlemont S, Lefebvre G, Duffner S, Garcia C. Class-balanced siamese neural networks. Neurocomputing 2018; 273: 47–56, https://doi.org/10.1016/j.neucom.2017.07.060.
6. Breunig M, Kröger P, Ng R, Sander J. LOF: Identifying Density-Based Local Outliers. 2000; 29: 104, https://doi.org/10.1145/342009.335388.
7. Canizo M, Triguero I, Conde A, Onieva E. Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. Neurocomputing 2019; 363: 246–260, https://doi.org/10.1016/j.neucom.2019.07.034.
8. Carvalho T P, Soares F A A M N, Vita R et al. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering 2019; 137: 106024, https://doi.org/10.1016/j.cie.2019.106024.
9. Castellani A, Schmitt S, Squartini S. Real-World Anomaly Detection by Using Digital Twin Systems and Weakly Supervised Learning.

IEEE Transactions on Industrial Informatics 2021; 17(7): 4733–4742, https://doi.org/10.1109/TII.2020.3019788.
10. Che C, Wang H, Fu Q, Ni X. Combining multiple deep learning algorithms for prognostic and health management of aircraft. Aerospace Science and Technology 2019; 94: 105423, https://doi.org/10.1016/j.ast.2019.105423.
11. Choi K, Yi J, Park C, Yoon S. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. IEEE Access 2021; 9: 120043–120065, https://doi.org/10.1109/ACCESS.2021.3107975.
12. Dong Y. Implementing Deep Learning for comprehensive aircraft icing and actuator/sensor fault detection/identification. Engineering Applications of Artificial Intelligence 2019; 83: 28–44, https://doi.org/10.1016/j.engappai.2019.04.010.
13. Dou S, Yang K, Poor H V. PC2A: Predicting Collective Contextual Anomalies via LSTM With Deep Generative Model. IEEE Internet of Things Journal 2019; 6(6): 9645–9655, https://doi.org/10.1109/JIOT.2019.2930202.
14. Ergen T, Kozat S S. Unsupervised Anomaly Detection With LSTM Neural Networks. IEEE Transactions on Neural Networks and Learning Systems 2020; 31(8): 3127–3141, https://doi.org/10.1109/TNNLS.2019.2935975.
15. Feng Z, Tang J, Dou Y, Wu G. Learning Discriminative Features for Semi-Supervised Anomaly Detection. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, IEEE: 2021: 2935–2939, https://doi.org/10.1109/ICASSP39728.2021.9414285.
16. Gers F A, Schmidhuber J, Cummins F. Learning to Forget: Continual Prediction with LSTM. Neural Computation 2000; 12(10): 2451–2471, https://doi.org/10.1162/089976600300015015.
17. Gradel S, Aigner B, Stumpf E. Model-based safety assessment for conceptual aircraft systems design. CEAS Aeronautical Journal 2022; 13(1): 281–294, https://doi.org/10.1007/s13272-021-00562-2.
18. Helbing G, Ritter M. Deep Learning for fault detection in wind turbines. Renewable and Sustainable Energy Reviews 2018; 98: 189–198, https://doi.org/10.1016/j.rser.2018.09.012.
19. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997; 9(8): 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
20. Hsu C-Y, Liu W-C. Multiple time-series convolutional neural network for fault detection and diagnosis and empirical study in semiconductor manufacturing. Journal of Intelligent Manufacturing 2021; 32(3): 823–836, https://doi.org/10.1007/s10845-020-01591-0.
21. Ince T, Kiranyaz S, Eren L et al. Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks. IEEE Transactions on Industrial Electronics 2016; 63(11): 7067–7075, https://doi.org/10.1109/TIE.2016.2582729.
22. Jalayer M, Orsenigo C, Vercellis C. Fault detection and diagnosis for rotating machinery: A model based on convolutional LSTM, Fast Fourier and continuous wavelet transforms. Computers in Industry 2021; 125: 103378, https://doi.org/10.1016/j.compind.2020.103378.
23. Jeyaraj A K, Liscouët-Hanke S. A Safety-Focused System Architecting Framework for the Conceptual Design of Aircraft Systems. Aerospace 2022; 9(12): 791, https://doi.org/10.3390/aerospace9120791.
24. Jiang F, Fu Y, Gupta B B et al. Deep Learning Based Multi-Channel Intelligent Attack Detection for Data Security. IEEE Transactions on Sustainable Computing 2020; 5(2): 204–212, https://doi.org/10.1109/TSUSC.2018.2793284.
25. Jiang G, He H, Xie P, Tang Y. Stacked Multilevel-Denoising Autoencoders: A New Representation Learning Approach for Wind Turbine Gearbox Fault Diagnosis. IEEE Transactions on Instrumentation and Measurement 2017; 66(9): 2391–2402, https://doi.org/10.1109/TIM.2017.2698738.
26. Jiang M, Hou C, Zheng A et al. Weakly Supervised Anomaly Detection: A Survey. 2023. doi:10.48550/arXiv.2302.04549, https://doi.org/10.48550/arXiv.2302.04549.
27. Kłosowski G, Rymarczyk T, Niderla K et al. Using an LSTM network to monitor industrial reactors using electrical capacitance and impedance tomography – a hybrid approach. Eksploatacja i Niezawodność – Maintenance and Reliability 2023. doi:10.17531/ein.2023.1.11, https://doi.org/10.17531/ein.2023.1.11.
28. Li G, Jung J J. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. Information Fusion 2023; 91: 93–102, https://doi.org/10.1016/j.inffus.2022.10.008.
29. Li X, Li J, Qu Y, He D. Semi-supervised gear fault diagnosis using raw vibration signal based on deep learning. Chinese Journal of Aeronautics 2020; 33(2): 418–426, https://doi.org/10.1016/j.cja.2019.04.018.
30. Liu C, Sun J, Liu H et al. Complex engineered system health indexes extraction using low frequency raw time-series data based on deep

learning methods. Measurement 2020; 161: 107890, https://doi.org/10.1016/j.measurement.2020.107890.

31. Liu F T, Ting K M, Zhou Z-H. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, 2008: 413–422, https://doi.org/10.1109/ICDM.2008.17.
32. Liu J, Song X, Zhou Y et al. Deep anomaly detection in packet payload. Neurocomputing 2022; 485: 205–218, https://doi.org/10.1016/j.neucom.2021.01.146.
33. Lopez Pinaya W H, Vieira S, Garcia-Dias R, Mechelli A. Chapter 11 - Autoencoders. In Mechelli A, Vieira S (eds): Machine Learning, Academic Press: 2020: 193–208, https://doi.org/10.1016/B978-0-12-815739-8.00011-0.
34. Mei S, Cheng J, He X et al. A Novel Weakly Supervised Ensemble Learning Framework for Automated Pixel-Wise Industry Anomaly Detection. IEEE Sensors Journal 2022; 22(2): 1560–1570, https://doi.org/10.1109/JSEN.2021.3131908.
35. Mitra S, Mukhopadhyay R, Chattopadhyay P. PSO driven designing of robust and computation efficient 1D-CNN architecture for transmission line fault detection. Expert Systems with Applications 2022; 210: 118178, https://doi.org/10.1016/j.eswa.2022.118178.
36. Moghaddam M, Chen Q, Deshmukh A V. A neuro-inspired computational model for adaptive fault diagnosis. Expert Systems with Applications 2020; 140: 112879, https://doi.org/10.1016/j.eswa.2019.112879.
37. Nguyen H D, Tran K P, Thomassey S, Hamad M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management 2021; 57: 102282, https://doi.org/10.1016/j.ijinfomgt.2020.102282.
38. Ning S, Sun J, Liu C, Yi Y. Applications of deep learning in big data analytics for aircraft complex system anomaly detection. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 2021; 235(5): 923–940, https://doi.org/10.1177/1748006X211001979.
39. Pang G, Shen C, van den Hengel A. Deep Anomaly Detection with Deviation Networks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage AK USA, ACM: 2019: 353–362, https://doi.org/10.1145/3292500.3330871.
40. Pang G, Shen C, Jin H, van den Hengel A. Deep Weakly-supervised Anomaly Detection. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, Association for Computing Machinery: 2023: 1795–1807, https://doi.org/10.1145/3580305.3599302.
41. Peterson L E. K-nearest neighbor. Scholarpedia 2009; 4(2): 1883, https://doi.org/10.4249/scholarpedia.1883.
42. Plakias S, Boutalis Y S. Fault detection and identification of rolling element bearings with Attentive Dense CNN. Neurocomputing 2020; 405: 208–217, https://doi.org/10.1016/j.neucom.2020.04.143.
43. Rai K, Hojatpanah F, Badrkhani Ajaei F, Grolinger K. Deep Learning for High-Impedance Fault Detection: Convolutional Autoencoders. Energies 2021; 14(12): 3623, https://doi.org/10.3390/en14123623.
44. Raman MR G, Somu N, Mathur A P. A multilayer perceptron model for anomaly detection in water treatment plants. International Journal of Critical Infrastructure Protection 2020; 31: 100393, https://doi.org/10.1016/j.ijcip.2020.100393.
45. Reddy K K, Sarkar S, Venugopalan V, Giering M. Anomaly Detection and Fault Disambiguation in Large Flight Data: A Multi-modal Deep Auto-encoder Approach. Annual Conference of the PHM Society 2016. doi:10.36001/phmconf.2016.v8i1.2549, https://doi.org/10.36001/phmconf.2016.v8i1.2549.
46. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 2015; 10(3): e0118432, https://doi.org/10.1371/journal.pone.0118432.
47. Schölkopf B, Platt J C, Shawe-Taylor J et al. Estimating the Support of a High-Dimensional Distribution. Neural Computation 2001; 13(7): 1443–1471, https://doi.org/10.1162/089976601750264965.
48. 48. Shen K, Zhao D. An EMD-LSTM Deep Learning Method for Aircraft Hydraulic System Fault Diagnosis under Different Environmental Noises. Aerospace 2023; 10(1): 55, https://doi.org/10.3390/aerospace10010055.
49. Shu X, Zhang S, Li Y, Chen M. An anomaly detection method based on random convolutional kernel and isolation forest for equipment state monitoring. Eksploatacja i Niezawodność – Maintenance and Reliability 2022; 24(4): 758–770, https://doi.org/10.17531/ein.2022.4.16.
50. Shyu M-L, Chen S-C, Sarinnapakorn K, Chang L. Principal Component-based Anomaly Detection Scheme. In Young Lin T, Ohsuga S,

Liau C-J, Hu X (eds): Foundations and Novel Approaches in Data Mining, Berlin/Heidelberg, Springer-Verlag: 2005; 9: 311–329, https://doi.org/10.1007/11539827_18.

51. Su S, Sun Y, Peng C, Wang Y. Aircraft Bleed Air System Fault Prediction based on Encoder-Decoder with Attention Mechanism. Eksploatacja i Niezawodność – Maintenance and Reliability 2023. doi:10.17531/ein/167792, https://doi.org/10.17531/ein/167792.
52. Sun J, Wang F, Ning S. Aircraft air conditioning system health state estimation and prediction for predictive maintenance. Chinese Journal of Aeronautics 2020; 33(3): 947–955, https://doi.org/10.1016/j.cja.2019.03.039.
53. Sun W, Shao S, Zhao R et al. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. Measurement 2016; 89: 171–178, https://doi.org/10.1016/j.measurement.2016.04.007.
54. Yang K, Ren J, Zhu Y, Zhang W. Active Learning for Wireless IoT Intrusion Detection. IEEE Wireless Communications 2018; 25(6): 19–25, https://doi.org/10.1109/MWC.2017.1800079.
55. Zaheer M Z, Mahmood A, Khan M H et al. An Anomaly Detection System via Moving Surveillance Robots with Human Collaboration. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, IEEE: 2021: 2595–2601, https://doi.org/10.1109/ICCVW54120.2021.00293.
56. Zeiser A, Özcan B, Van Stein B, Bäck T. Evaluation of deep unsupervised anomaly detection methods with a data-centric approach for on-line inspection. Computers in Industry 2023; 146: 103852, https://doi.org/10.1016/j.compind.2023.103852.
57. Zhang C, Song D, Chen Y et al. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. Proceedings of the AAAI Conference on Artificial Intelligence 2019; 33(01): 1409–1416, https://doi.org/10.1609/aaai.v33i01.33011409.
58. Zhang J, Sun Y, Guo L et al. A new bearing fault diagnosis method based on modified convolutional neural networks. Chinese Journal of Aeronautics 2020; 33(2): 439–447, https://doi.org/10.1016/j.cja.2019.07.011.
59. Zhao G, Zhang G, Ge Q, Liu X. Research advances in fault diagnosis and prognostic based on deep learning. 2016 Prognostics and System Health Management Conference (PHM-Chengdu), 2016: 1–6, https://doi.org/10.1109/PHM.2016.7819786.
60. Zhao Y, Nasrullah Z, Li Z. PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of Machine Learning Research 2019; 20(96): 1–7.
61. Zhao Y, Zhao H, Ai J, Dong Y. Robust Data-Driven Fault Detection: An Application to Aircraft Air Data Sensors. International Journal of Aerospace Engineering 2022; 2022: 1–17, https://doi.org/10.1155/2022/2918458.
62. Zhi Z, Liu L, Liu D, Hu C. Fault Detection of the Harmonic Reducer Based on CNN-LSTM With a Novel Denoising Algorithm. IEEE Sensors Journal 2022; 22(3): 2572–2581, https://doi.org/10.1109/JSEN.2021.3137992.
63. Zhou Y, Song X, Zhang Y et al. Feature Encoding With Autoencoders for Weakly Supervised Anomaly Detection. IEEE Transactions on Neural Networks and Learning Systems 2022; 33(6): 2454–2465, https://doi.org/10.1109/TNNLS.2021.3086137.
64. Zhou Z-H. A brief introduction to weakly supervised learning. National Science Review 2018; 5(1): 44–53, https://doi.org/10.1093/nsr/nwx106.