Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07903-0

ORIGINAL ARTICLE

Few-shot learning for modeling cyber physical systems in non-stationary environments

Stavros Ntalampiras[1] · Ilyas Potamitis[2]

Received: 28 June 2022 / Accepted: 28 September 2022

Abstract
This paper proposes a modeling scheme for cyber physical systems operating in non-stationary, small data environments. Unlike the traditional modeling logic, we introduce the few-shot learning paradigm, the operation of which is based on quantifying both similarities and dissimilarities. As such, we designed a suitable change detection mechanism able to reveal previously unknown operational states, which are incorporated in the dictionary online. We elaborate on spectrograms extracted from high-resolution ultrasound depth sensor timeseries, while the backbone of the proposed method is a Siamese Neural Network. The experimental scenario considers data representing liquid containers for fuel/water when the following five operational states are present: *normal*, *accident*, *breakdown*, *sabotage*, and *cyber-attack*. Thorough experiments were carried out assessing every aspect of the present framework and demonstrating its efficacy even when very few samples per class are available. In addition, we propose a probabilistic data selection scheme facilitating one-shot learning. Last but not least, responding to the wide requirement for interpretable AI, we explain the obtained predictions by examining the layer-wise activation maps.

Keywords Cyber-physical systems · Cybersecurity · Few-shot learning · Deep learning · Online learning · Fault diagnosis · Cyber-attacks

1 Introduction

The intersection between the scientific fields of artificial intelligence and more specifically machine learning with Cyber Physical Systems (CPS) is receiving ever-increasing attention by the community [1–3]. Given that the cyber layer has been introduced to a vast gamut of systems, including critical infrastructures, Internet of Things [4], etc., manual inspection of the quality of the communicated information became impossible in practise, thus automatising cybersecurity mechanisms comprises a necessity of the utmost urgency. Unfortunately, the operation of CPSs may be negatively affected by a great range of conditions including but not limited to sensor faults, state drifts, cyber-attacks [5], environmental changes, time-variances, etc. At the same time, one has to consider that the large-scale of CPSs as well as the existence of potential interconnections which heavily burden the construction of analytical models explaining the operation of interconnected CPSs [6, 7]. As such, cybersecurity analysts process the available to data to create models representing the data-generating process. In this direction, AI-based tools and methodologies are able to detect and analyze irregularities in the acquired data, hence potentially revealing the existence of system faults, cyber-attacks [8], etc.

The related literature includes a plethora of methodologies which basically follow the same principal pipeline where parameters characteristic of the problem at hand are extracted and subsequently modeled using generative (e.g., hidden Markov models) or discriminative machine learning models (e.g., support vector machines [9], deep neural

✉ Stavros Ntalampiras
stavros.ntalampiras@unimi.it

Ilyas Potamitis
potamitis@hmu.gr

[1] Department of Computer Science, University of Milan, via Celoria 18, 20133 Milan, Italy

[2] Department of Music Technology and Acoustics, Hellenic Mediterranean University, Gianni Kornarou, Estavromenos 1, 71410 Iraklio, Greece

Published online: 15 October 2022

networks [10, 11], etc.). Several strong assumptions are made during the specific modeling process:

(a) rich (or at least substantial) data availability with respect to every considered class,
(b) *a-priori* knowledge of the class dictionary, and
(c) availability of reliable domain expert knowledge for feature engineering.

The majority of existing works typically train and evaluate the designed solutions within a closed-world setting, i.e., assuming that train and test data belong to the same distributions. However, this does not represent well real-world conditions, where one has to deal with non-stationary and open environments [12]. At the same time, there could be biases hidden inside the data making the produced model favoring certain patterns and/or types of predictions [13].

This work argues that the above-mentioned hypotheses are quite strong leading to systems which are not directly applicable to real-world CPSs, where

(a) it is unrealistic to assume complete knowledge of the class dictionary since new classes of faults, attacks, etc. may appear at any point in time,
(b) furthermore, we cannot assume availability of an amount of data adequate to train deep models, or at least, that is not true for part of the classes, e.g., rarely occurring faults, cyber-attacks which can have catastrophic consequences,
(c) as such, it is strong to assume that domain experts would know the important characteristics of newly appearing states in order to engineer descriptive features.

Keeping the above-mentioned requirements in mind, we propose to suitably enhance the one-shot learning paradigm [14, 15] to the present problem, where the main limitation is the fact that we may observe only a handful of examples during model training. More specifically, recognition is carried out via a model learning to assess similarities between novel data and those available during training. As such, the proposed paradigm is radically different than the existing line of thought, where the solutions seek to identify hyperplanes separating classes (discriminative modeling) or building representations estimating class distributions (generative modeling). To the best of our knowledge, such a solution has never been explored in the CPS research domain.

The two main modules of the proposed solution are *change detection*, where we discover previously unseen CPS states and *state identification*, where the algorithm identifies the current operational state. The first one detects a new state in case the observed data are labeled as dissimilar to every known state, while the second assigns the state with the highest similarity score to the observed data. Without loss of generality, we operate on a dataset of limited dimensions [16] including data of a CPS consisting of liquid containers for fuel or water, along with its automated control and data acquisition infrastructure. We elaborate on high-resolution ultrasound depth sensor data, which is representative of the differences existing between normal and anomalous data. Toward eliminating the need for domain expert knowledge we propose a standardize feature set, i.e., spectrograms characterizing the available operational states. Subsequently, we train a Siamese Neural Network (SNN) on learning relationships between spectrograms coming from same or different CPS states. We thoroughly assess the performance of the proposed system using appropriate figures of merit in (a) identifying CPS operational states, (b) detecting new ones, (c) incorporate them in the class dictionary, (d) operate in non-stationary environments. Toward relaxing further data quantity requirements, we designed a data selection mechanism estimating the distributions of the available samples using Gaussian Mixture models. By considering intra- and inter-class Kullback-Leibler-based distances, the proposed algorithm identifies a unique sample to represent an operational state, which is used to learn the SNN in one-shot mode. Finally, we provide an interpretation of the obtained results, which is a demand of the utmost importance for developed AI-based tools and methodologies [17], via analyzing the activation maps.

In the following, we (a) formalize the problem, (b) delineate the proposed solution, (c) describe the experimental protocol along with a detailed analysis of the obtained results, (d) draw conclusions and briefly discuss potential extensions.

2 Problem formulation

We assume availability of data characterizing operational states of cyber-physical systems, i.e., a labeled training set $\mathcal{TS}$. These states form a dictionary $\mathcal{D} = \{S_1, S_2, \ldots, S_n, \}$, where S_i denotes the i-th state and n the number of known states during training. They follow a consistent, yet unknown probability density function $P_i, 1 < i < n$ [18]. On the contrary, no assumption is made regarding the composition of $\mathcal{D}$, i.e., it may encompass nominal conditions, component faults, cyber-attacks, drifts, etc. Aiming at representing real-world conditions, we drastically restrict the number of available samples per state [16]. On top of that, the cardinality of $\mathcal{D}$ is known only up to a certain extent, i.e., previously unseen operational states may appear at any point in time. The overall goal is to identify the operational state, promptly detect changes in composition and/or size of $\mathcal{D}$ as well as incorporate such changes online.

3 Few-shot learning for identification of operational states

The proposed solution encompasses a support set of labeled examples representing the known operational states denoted as $\mathcal{S}$ and an SNN learning *similar* and *dissimilar* relationships of the classes in $\mathcal{TS}$. The overall block diagram is depicted in Fig. 1 where we observe that the system receives two inputs (spectrograms of operational states) and processes them using a symmetrical network architecture ending at a common point where a prediction is made based on the maximum similarity/dissimilarity score.

The design of the proposed solution is described in the next subsections as follows:

(a) SNN design, architecture and learning,
(b) feature extraction process, and
(c) operational state identification and change detection algorithm.

3.1 Siamese neural networks

The SNN is composed of a twin network each on processing a different input, while their outputs are connected and terminate to a common point [19] (see Fig. 1). In the ending point, the SNN calculates the distance between the two output representations as they produced by each network using predetermined distance metric. At first, spectrograms representing operational states of the considered CPS are extracted and fed to each network. As we see in Fig. 1, each network processes the input spectrogram interdependently from the other without any type of connection. However, they attempt to satisfy the same optimization function and as such, the learned weights are linked and produce representations which are closely-located representations in the feature space. On top of that, the specific SNN architecture encodes a learning process rendering it exchangeable, i.e., if the networks/inputs were to be reversed (top/bottom), the output distance metric would lead to the same value. It should be noted that the proposed SNN incorporates binary cross-entropy loss followed by a sigmoid activation during distance assessment.

Having designed the twin architecture, the next step is focused on forming the structure of each network. Lately, Convolutional Neural Networks (CNNs) have provided excellent performance in audio signal processing systems including a great variety of tasks such as environmental sound recognition [20], music information retrieval [21]. Hence, we decided to populate each SNN with a series of convolutional layers, the number of which is determined during the model optimization phase.

Interestingly, CNNs consist of a series of stacked layers, where convolutions are succeeded by max-pooling operations. Such processing emphasizes localized patterns in the 2D plane, while each hidden unit accesses only a limited part of the input, the so-called *receptive field*. Thus the network is able to encode specific spectrogram regions, which may be distinctive and assist in assessing similarities and dissimilaties existing between the pair of inputs. Interestingly, dimensionality of the learned weights is suitably controlled by max-pooling layers which robustify the network to translational shifts [20], i.e., structural deviations in the input data are compensated by the included max-pooling operations.

Moreover, we employed rectified linear units (ReLU), i.e., the activation function is $f(x) = max(0, x)$. The specific choice is motivated by their superiority over traditional units, e.g., logistic sigmoid and hyperbolic tangent

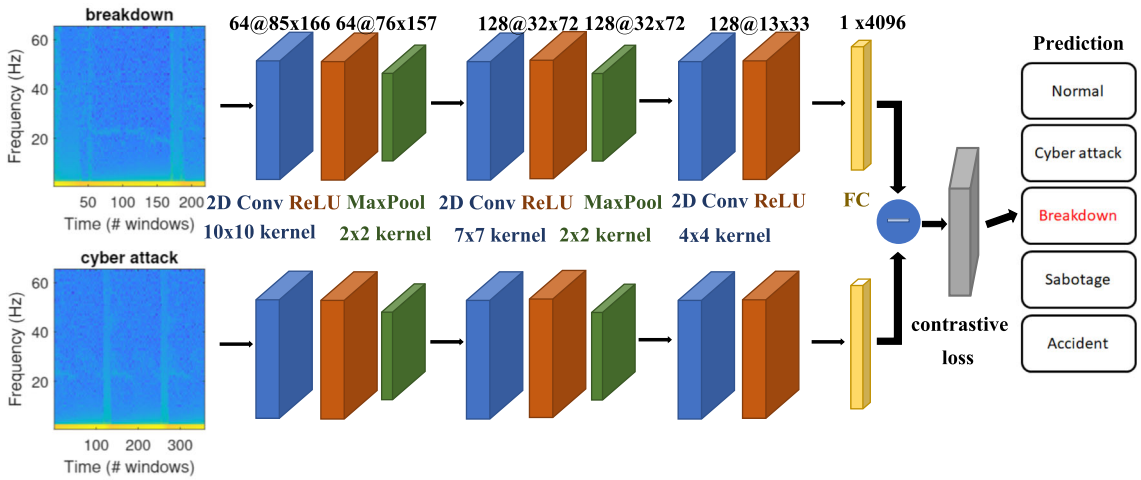

Fig. 1 The pipeline of the proposed one-shot learning scheme using Siamese neural networks. Each input is passed though a series of convolutional, ReLU and max-pooling layers completed by a common end based on binary cross-entropy loss

as gradient propagation does not suffer from saturations effects, they are biologically possible and sparse activation organization [20]. Regardless of their simplicity, neural networks with such activation functions demonstrate substantial discriminatory properties.

3.2 SNN architecture and learning

Following the optimization process, as shown in Fig. 1, each SNN twin is composed of three convolutional layers, where the initial two are followed by ReLU and max-pooling ones. The concluding layer is a fully-connected one which flattens the so-far result and include the final input representation. The proposed SNN is completed by a distance operation, namely binary cross-entropy loss, which is succeeded by a fully-connected layer and a sigmoid function assessing similarity between input pair.

Going into the parameterization of the presented neural architecture, the convolutional filters have a stride equal to 1 and kernels as shown in Fig. 1, while max-pooling layers have 2×2 kernels with $stride = 2$. The employed learning process targets the minimization of binary cross-entropy loss among network's prediction and ground truth using the standard version of backpropagation algorithm. Minibatch size is chosen according to the $\mathcal{TS}$ size at a learning rate of 6e−5. Weight initialization is carried out via narrow normal distributions with zero-mean and 0.01 standard deviation. Last but not least, the maximum number of permitted iterations is 2000.

3.3 Feature extraction

We elaborate on ultrasound depth sensor data, which are characterized by high resolution and as such, highlighting the discrepancies between normal and anomalous data. Aiming at eliminating the feature engineering process, we divide the signal into frames of 128 samples ovelapping by 100 samples using a Hamming window and compute the spectrogram with an FFT size equal to 128. Spectrograms associated with the five operational states considered in this work are illustrated in Fig. 2. We observe that lower frequency parts are associated with higher energy values for every operational state. However, the frequency content exhibits differences across states and as such, it could be informative for classification purposes. More specifically, we observe that accidents exhibit high energy content in a discrete but homogeneous way across frequency bands. At the same time, the energy of breakdowns in higher bands is not as siginifcant similar to the cyber attack state which demonstrates such behavior in shorter time intervals. Normal state starts with low energy content for the majority of frequency bands, while sabotage is the most distinctive state as it is characterized by high energy across both frequency and time dimensions.

3.4 Identification of operational state and change detection

The proposed SNN, illustrated in Fig. 1 learns to identify similar and dissimilar pairs of input spectrograms. Keeping in mind the requirements outlined in Sect. 1, we developed a straightforward extension suitable for change detection. After contrasting the unknown input with every class existing in set $\mathcal{S}$ and dictionary $\mathcal{D}$, a change is flagged in case the novel spectrogram is recognized as dissimilar to every available class. Thus, we form an additional class and appropriately augment $\mathcal{S}$ and $\mathcal{D}$ using the specific spectrogram. Interestingly, SNN can successfully address classification tasks in poor data environments [22].

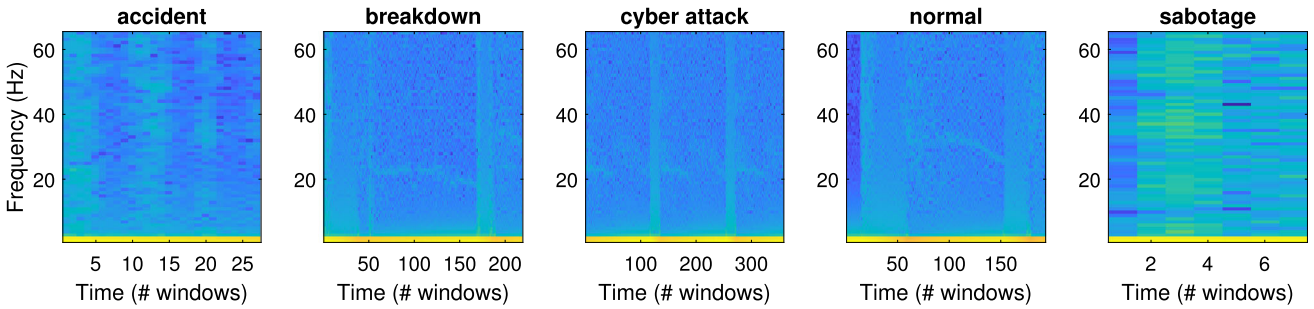

Fig. 2 Spectrograms representing the considered operational states, i.e., *accident*, *breakdown*, *cyber attack*, *normal*, and *sabotage*

Algorithm 1 The proposed operational state identification algorithm using ultrasound signals based on few-shot learning ($\lfloor\bullet\rceil$ denotes the cardinality operator).

1: Input: test data t, trained SNN $\mathcal{N}$, dictionary $\mathcal{D}$, where each class is represented by extracted spectrograms of the support set $\langle\mathcal{S}_{i=1}^{i=d}\rangle$, where $d = \lfloor\mathcal{D}\rceil$.
2: Extract spectrogram s of t;
3: Initialize similarity vector $V = []$;
4: **for** j=1 : $\lfloor\mathcal{D}\rceil$ **do**
5: **for** i=1 : $\lfloor\mathcal{S}\rceil$ **do**
6: Query $\mathcal{N}$ with the pair $\{s, \mathcal{S}_j^i\}$ and get similarity score $V(j, i)$;
7: **end for**
8: **end for**
9: Predict the class maximizing similarity score $S^* = \arg\max_{\mathcal{S}}\{V(:, i)\}$ and assign it to t;

On the opposite case, when the unknown example is predicted as similar to one or more classes, the one with the highest similarity is selected. The proposed operational state prediction algorithm, illustrated in Alg. 1, necessitates as inputs

- the test data t to be used for feature extraction,
- the trained SNN $\mathcal{N}$, and
- the dictionary $\mathcal{D}$, where each class is represented by extracted spectrograms of the support set $\langle\mathcal{S}_{i=1}^{i=d}\rangle$.

Subsequently, it extracts the spectrogram s of the unknown example t using the same process outlined in Sect. 3.3 (Alg. 1, line 2) and initializes similarity vector V (Alg. 1, line 3). Afterward, it queries $\mathcal{N}$ using the existing pair combinations which outputs the corresponding similarity scores and updates V (Alg. 1, line 4-8). The support set, i.e., the known samples are the ones populating $\mathcal{TS}$ and the final score is normalized by the number of available samples per class. The last step of the algorithm assigns to t the label of the class maximizing the similarity score in V (Alg. 1, line 9). Importantly, such an Algorithm comprises a common framework able to process data which may belong to any operational state including both cyber attacks and faulty states.

3.5 Probabilistic data selection for one-shot learning

To further minimize the required data quantity, we designed a scheme for selecting solely one sample to represent each class, thus realizing one-shot learning [23]. Keeping in mind that the proposed methodology learns to assess similarities and dissimilarities, each class is represented by the sample which satisfies a twofold criterion, i.e.,

- minimizing the sum of distances to intra-class samples, and
- maximizing the sum of distances to inter-class samples.

To this end, we defined a suitable distance metric. Starting from the extracted spectrograms, Gaussian Mixture models (GMM) are used to estimate their distributions. As such, we move from the feature space to the probabilistic plane which may provide improved generalization of the represented classes over novel samples.

Let $\mathcal{G}^s$ characterized by set of vectors $\{\mu^s, \sigma^s\}$ denote the GMM approximating the distribution of the spectrogram representing the operational state s. In order to position the available data samples expressed in GMMs in the probabilistic plane, we suitably adapted the Kullback-Leibler Divergence (KLD). The KLD between two n-dimensional probability distributions S and N is defined as [24]:

$$KL(S\|N) = \int_{R^n} p(X,S) log \frac{p(X,S)}{p(X,N)} dx \tag{1}$$

Even though KLD is able to quantify the distance existing between two probability distributions, in its current form, it cannot be considered as a distance metric since it does not satisfy the property of symmetry [25]. Thus, we employed its symmetric form given by the following formula

$$KL_d(S\|N) = D(S\|N) + D(N\|S). \tag{2}$$

Moreover, when S and N are in the form of GMMs, KL_d becomes

$$KL_d(S\|N) = \int_{R^n} S(x) log \frac{N(x)}{S(x)} dx. \quad (3)$$

To the best of our knowledge, a closed-form solution for Eq. 3 does not exist, hence we rely on the empirical mean, i.e.,

$$KL_d(S\|N) \approx \frac{1}{m} \sum_{i=1}^{m} log \frac{N(x_i)}{S(x_i)} \quad (4)$$

under the assumption that the number of Monte Carlo draws m is sufficiently large. It should be noted that during our experiments we set $m = 5000$.

Based on the distance metric defined in Eq. 4, we calculate the intraclass sum of distances and the corresponding interclass sum for every available sample $i \in S$ as follows:

$$\mathcal{D}_i^r = \sum_{j \in S, i \neq j, j=1}^{j=\lfloor S \rceil} KL_d(\mathcal{G}^i \| \mathcal{G}^j) \quad (5)$$

$$\mathcal{D}_i^a = \sum_{j \notin S, j=1}^{i=\lfloor \mathcal{TS} \rceil} KL_d(\mathcal{G}^i \| \mathcal{G}^j) \quad (6)$$

Finally, for each operational state, we choose the samples minimizing the quantity $\mathcal{D}^r - \mathcal{D}^a$ to learn the SNN in one-shot mode. The same samples populate the support set as well. The proposed probabilistic data section scheme is illustrated in Fig. 3.

4 Experimental set-up and results

This section describes the experimental set-up and analyzes the obtained results. It is organized as follows: (a) (b) employed dataset, (c) suitably-formed figures of merit, (d) contrasted method, (e) obtained results, and (f) interpretation of SNN's decision making process. It should be noted that we addressed both the binary (normal vs. abnormal) as well as the full-range five class classification problem.

4.1 Dataset

The employed dataset was designed for studying anomalies and malicious acts in CPSs [16]. It represents the operation of liquid containers for fuel/water, along with its automated control and data acquisition infrastructure. Conveniently, the dataset is publicly available for research purposes facilitating reproducibility and comparison between different solutions. The included temporal series are representative of five operational scenarios, i.e., *normal*, *accident*, *breakdown*, *sabotage*, and *cyber-attack* corresponding to 15 different real situations. There are 2-6 examples per class which fits well the problem specifications analyzed in Sects. 1 and 2. We elaborate on high-resolution ultrasound depth sensor data, which is representative of the differences existing among the various operational states. These are divided into frames of 128 samples overlapping 100, while the FFT size was 128. The interested reader is referred to [16] for more information. The specific dataset fits well the aim of this research as it satisfies the small data requirement, while including a wide range of abnormal operational states which are typically treated independently in the related literature [26].

4.2 Figures of merit and contrasted approach

In thoroughly assessing the capabilities of the designed systems we employed standardized figures of merit facilitating comparability with some target approaches. Interestingly, within the few-shot learning paradigm we can derive confusion matrices evaluating similarities and dissimilarities. To this end, the following matrix was defined:

$$\mathcal{M}^s = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \quad (7)$$

where

- s_{xx} (in %) denotes the number of times that spectrograms fed in the x input of SNN were identified as similar to spectrograms coming from the same class,

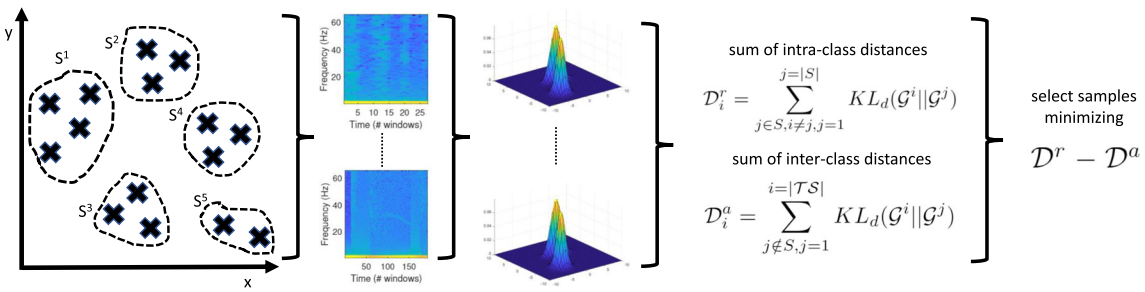

Fig. 3 Probabilistic data selection for one-shot learning

- s_{xy} (in %) denotes the number of times that spectrograms fed in the x input of SNN were identified as dissimilar to spectrograms coming from the same class,

Evidently, the objective is to maximize the values in the diagonal. A matrix assessing the dissimilarities $\mathcal{M}^d$ can be defined in an analogous way where we aim at minimizing its diagonal. It should be mentioned that the sum of similarity and dissimilarity matrices characterizing the accuracy of a given method is 100%, i.e., $\mathcal{M}^s + \mathcal{M}^d = 100$ for every element [27].

The proposed method is compared to the k-NN algorithm as, to the best of our knowledge, is the only alternative method able to operate under such restrictive assumptions.

4.3 Results

The performance of the proposed solution was evaluated extensively from different points of view. At first, we tested the behavior when knowledge regarding composition and size of $\mathcal{D}$ is unknown, i.e., a limited number of states is known during training. We considered the following pairs of known-unknown classes $\{(2,3),(3,2),(4,1),(5,0)\}$ while they were chosen randomly. It should be noted that the minimum number of classes allowing learning similar and dissimilar relationships is two, which comprises the minimum amount of classes that is assumed to be known during training. Such an assumption is not restrictive for the majority of CPS applications where typically data representing more than two classes are available. The experiment corresponding to each class setting was iterated 100 times and the results were averaged. Fig. 4 illustrates the mean and standard deviation of the obtained recognition rates. During this process, model optimization and learning were carried out using half of the available dataset, while testing on the rest. It should be noted that similar and dissimilar input pairs were produced randomly.

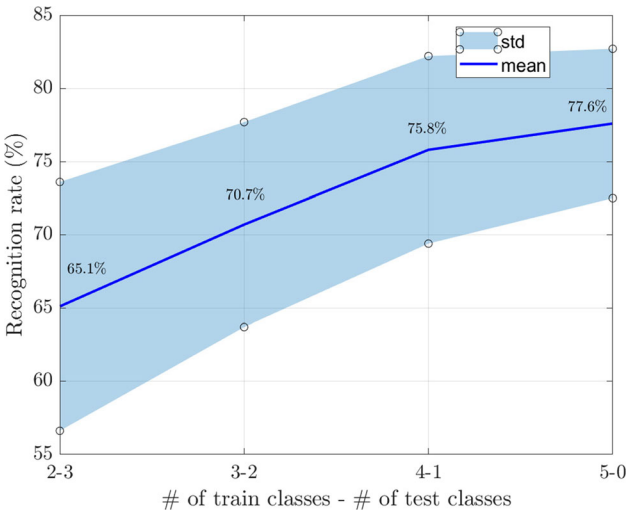

Fig. 4 Recognition rates achieved by the proposed SNN while varying the number of known classes during the model learning process

We observe that the recognition rates reached by the proposed system range from 65.1% in the (2,3) setting to 77.6% in the (5,0) setting. On top of that, standard deviation decreases as data representing more classes become available, i.e., from 8.5% to 5.1%. As expected, the performance of the proposed system improves as the amount of classes existing in $\mathcal{TS}$ increases. Interestingly, the SNN is not only able to operate in a small data environment but the achieved rates are promising. We infer that transforming the classification problem to a similarity one is particularly relevant in identifying every operation state, i.e., normal, accident, breakdown, sabotage, and cyber-attack. Even when only two classes are included in $\mathcal{TS}$, the achieved recognition rate is significantly higher than chance (20%). As expected, the rate increases as more data become available since it contributes toward similarity and dissimilarity learning. Importantly, when every class is considered to be *a-priori* known, the performance is more than satisfactory given the low amount of available data. In the specific (5,0) scenario, euclidean distance-based k-NN provided a recognition rate of 54.7% underlining the superiority of the proposed relationship-based system. In fact, the proposed Siamese network is able to significantly outperform the k-NN based solution in every considered class setting. Unfortunately, comparing other machine learning-based solutions, support vector machines, artificial neural networks, hidden Markov models, etc. is not feasible due to their tendency to overfit when so few data are available during training [28].

The confusion matrix $\mathcal{M}^s$ obtained in the (5,0) setting is presented in Table 1. We can see that the state recognized with the highest rate is the cyber-attack (91.4%), while the

Table 1 $\mathcal{M}^s$ (in %) achieved by SNN trained and optimized on 50% of samples per class and tested on the remaining ones in the (5,0) class setting (maximum rates are emboldened). Average recognition accuracy is 77.6%

Input 1	Input 2				
	Accident	Breakdown	Attack	Normal	Sabotage
Accident	**88.1**	1.7	-	6.8	3.4
Breakdown	–	**52.4**	19.6	–	28
Attack	–	8.6	**91.4**	–	–
Normal	3.5	7	–	**86**	3.5
Sabotage	14.8	3.7	–	11.1	**70.4**

one with the lowest is breakdown (52.4%). Such a behavior is directly related with the intra-class similarity and inter-class dissimilarity characterizing the specific classes. Cyber-attacks tend to exhibit quite different spectral patterns with respect to the rest of the classes. Breakdown class exhibits similarities with cyber-attacks and sabotage, thus the great amount of misclassifications. Importantly, miclassifications with the normal operational state are limited, hence the proposed solution may serve anomaly detection tasks as explained next. Table 2 evaluates relationship learning in the (5,0) scenario. There, we see that the SNN learns the similar relationships (86.2%) better than the dissimilar ones (69.7%) with an average recognition rate equal to 78%. As such, the identification capabilities exhibited so far are based more on the learned intra-class similarities.

In the next phase, we evaluated a simplified version of the present problem which may consist the first line of defense in monitoring CPSs. We experimented with the two-class problem, i.e., normal vs. abnormal operational states, where abnormal includes accident, breakdown, sabotage, and cyber-attack. The obtained similarity matrix $\mathcal{M}^s$ is presented in Table 3. As expected, we see that the recognition rates increase substantially reaching 96.2% for similar and 92.8% for dissimilar relationships. We argue that the present learning framework can address the simplified problem quite efficiently. That is confirmed by the results included in the confusion matrix presented in Table 4 where the average recognition rate for normal and abnormal states is 95.6%. On the contrary, the k-NN based solution reached 64.7%.

4.4 Evaluation of the system learnt with one sample

In this section, we report the results obtained after the application of the data selection algorithm outlined in Sect. 3.5. During the parameterization phase, we experimented various number of Gaussian components to estimate the distribution of each available sample. The explored number of Gaussian components comes from the following set: $\{2, 4, 8, 16\}$ while, during cluster initialization, the maximum permitted number of k-means iterations was set to 50.

Thus, the system was trained on one sample per class and evaluated on the rest of the dataset. The support set also includes one sample per class. The obtained accuracy on the full-range 5 class problem was equal to 55.9%, while the rate on the 2-class problem was 78.2%. The specific scheme outperformed random data selection, which provided 37.6% and 54.9%, respectively. Interestingly is slightly outperformed the k-NN based solution as well. That said, the achieved rates are significantly lower than the corresponding ones exploiting more training data as presented above. It comes out that the SNN trained on one sample per class is not able to generalize well over the test dataset meaning that information included in greater amount of data is required to address the task at hand.

Table 2 Similarities-dissimilarities confusion matrix (%) in the (5,0) setting obtained with SNN trained on 50% of the available data. The average recognition rate is 78%

Presented	Predicted	
	Similar	Dissimilar
Similar	**86.2**	13.8
Dissimilar	30.3	**69.7**

The maximum rates are emboldened

Table 3 Similarities-dissimilarities confusion matrix (%) in the two-class setting obtained with SNN trained on 50% of the available data. The average recognition rate is 94.5%

Presented	Predicted	
	Similar	Dissimilar
Similar	**96.2**	3.8
Dissimilar	7.2	**92.8**

The maximum rates are emboldened

Table 4 $\mathcal{M}^s$ (in %) achieved by SNN in the 2-class scenario. Average recognition accuracy is 95.6%

Presented	Predicted	
	Normal	Abnormal
Normal	94.5	5.5
Abnormal	3.3	96.6

4.5 Activation maps

This experimental phase examines the way SNN processes the spectrograms by means of the considered convolutional layers emphasizing on the regions employed to assess similar/dissimilar relationships. To this end, we visualized the parts of the spectrogram which activated the network layers as the input advances through them triggering the included algebraic operations.

Such activations maps representing the relevant regions of samples belonging to every considered class are demonstrated in Fig. 5. The maps show the evolution of the activations as the spectrogram propagates through every convolutional layer.

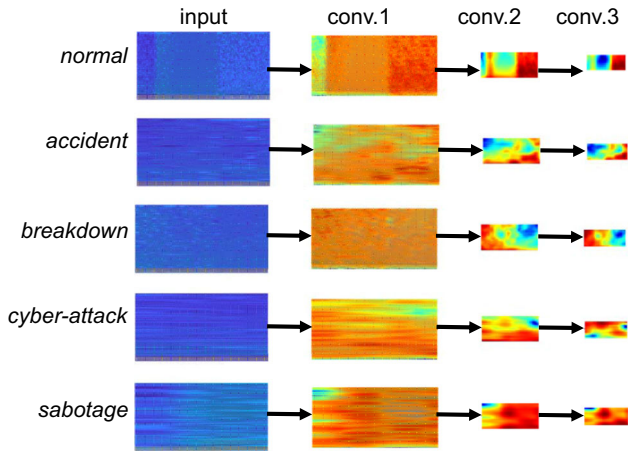

Fig. 5 Activation maps obtained from every convolutional layer when the trained SNN processes samples coming from every operational state

Each convolutional layer simplifies the representation extracted from the previous one, while localizing characteristic spectrogram regions useful for assessing similar/dissimilar relationships. It is evident that not every part of the spectrogram is equally distinctive for every state. We observe that SNN assigns different levels of significance on the spectrogram content based on the operational state undergoing processing. More precisely,

- *normal state*: most of the emphasis is placed on the lower frequencies, followed in time by low-significance content,
- *accident state*: it is identified by very low frequency and narrow content, while the higher frequencies are considered only partially,
- *breakdown state*: it is recognized by early and mostly high frequency content,
- *cyber-attack state*: continuous high-frequency content plays the most important role as regards to this state, and
- *sabotage state*: processing here is based on the use of a wide part of the spectrum confirming the high intra-class diversity.

A thorough analysis of the SNN's operation explaining its final prediction may provide a meaningful interpretation, which constitutes a strong requirement towards robust, verifiable and trustworthy machine learning based solutions and a wider acceptance of such solutions [17, 29, 30].

5 Conclusion and future work

This work presented a novel solution for the automatic identification of CPS operational states relaxing a series of strong assumptions made in the related literature. We considered data representing the operation of liquid containers for fuel/water, along with its automated control and data acquisition infrastructure. Interestingly, the proposed solution is able to operate in non-stationary environments where state dictionary $\mathcal{D}$ is only partially known. To this end, the system relies on a suitably-designed change-detection mechanism able to reveal new classes and incorporate them in $\mathcal{D}$. At the same time, the solution operates efficiently in a small data environment since unbiased data characterizing the entire range of classes representing the task at hand is quite limited. The few-shot learning based solution was contrasted with k-NN, confirming its superiority. Finally, SNN's predictions are interpretable by examining the activation maps of the convolutional layers, which are perceptible by humans. Importantly, we outlined the design of mechanism based on probabilistic distances facilitating one-shot learning. We argue that a significant part contributing to the success of this solution is its ability to simultaneously consider both similarities and dissimilarities to known operational states.

Few-shot learning not only offers superior to the k-NN performance but, at the same time, we obtain an actual model learning similar and dissimilar relationships existing in the training data. In addition, the extracted interpretations of the decisions made by the systems in terms of feature space importance (see sec. 4.5) provide interesting insights as to which feature parts are relevant to uniquely characterize each operational state.

The recently presented report by the Capgemini group in [31] highlights the popularity of AI-based tools in Cybersecurity as threats overwhelm cyber analysts who fail to keep pace with the ever-increasing types of attacks. Thus, urgent requirements for such tools and methodologies include the use of small data, consider non-stationary environments, end-to-end approaches where the need for domain expert knowledge is minimized, and interpretable predictions. The proposed few-shot learning system responds to every requirement since (a) it requires a restricted amount of training data, (b) it is able to incorporate non-stationarities on-the-fly, (c) it does not require a significant level of domain expertise, (d) explains the predictions regarding operational states, and (e) it is flexible and can adapt to other Cybersecurity tasks of similar requirements with minor modifications.

Our future works include:

(a) adaptation of the few-shot learning paradigm to different problems of similar requirements,
(b) experimenting and formulating sufficient conditions as regards to dataset composition and quantity in order to boost the achieved performance,

(c) extend the present framework towards considering data belonging to diverse modalities which may provide improved performance [32],
(d) addition of auditability, i.e., the operator should be able to "open" and check the internal state of the deployed system at any point in time and, especially when a prediction is carried out.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement.

Data availability The dataset used in this research is publicly available at https://doi.org/10.1016/j.dib.2017.07.038.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Samtani S, Kantarcioglu M, Chen H (2020) Trailblazing the artificial intelligence for cybersecurity discipline. ACM Trans Manage Inf Syst 11(4):1–19. https://doi.org/10.1145/3430360
2. Shaukat K, Luo S, Varadharajan V, Hameed IA, Xu M (2020) A survey on machine learning techniques for cyber security in the last decade. IEEE Access 8:222310–222354. https://doi.org/10.1109/access.2020.3041951
3. Shone N, Ngoc TN, Phai VD, Shi Q (2018) A deep learning approach to network intrusion detection. IEEE Trans Emerg Top Comput Intell 2(1):41–50. https://doi.org/10.1109/TETCI.2017.2772792
4. Mosenia A, Jha NK (2017) A comprehensive study of security of internet-of-things. IEEE Trans Emerg Top Comput 5(4):586–602. https://doi.org/10.1109/TETC.2016.2606384
5. Liu C, Cronin P, Yang C (2020) Securing cyber-physical systems from hardware trojan collusion. IEEE Trans Emerg Top Comput 8(3):655–667. https://doi.org/10.1109/TETC.2017.2787694
6. Alippi C, Ntalampiras S, Roveri M (2017) Model-free fault detection and isolation in large-scale cyber-physical systems. IEEE Trans Emerg Top Comput Intell 1(1):61–71. https://doi.org/10.1109/TETCI.2016.2641452
7. Wan X, Han T, An J, Wu M (2021) Fault diagnosis for networked switched systems: An improved dynamic event-based scheme. IEEE Trans Cyber, pp 1–12. https://doi.org/10.1109/TCYB.2021.3049838
8. Huang X, Dong J (2018) Reliable control policy of cyber-physical systems against a class of frequency-constrained sensor and actuator attacks. IEEE Trans Cyber 48(12):3432–3439. https://doi.org/10.1109/TCYB.2018.2815758
9. Aburomman AA, Reaz MBI (2016) A novel SVM-kNN-PSO ensemble method for intrusion detection system. Appl Soft Comput 38:360–372. https://doi.org/10.1016/j.asoc.2015.10.011
10. Gümüşbaş D, Yıldırım T, Genovese A, Scotti F (2020) A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Syst J, pp 1–15. https://doi.org/10.1109/JSYST.2020.2992966
11. Coulter R, Han Q-L, Pan L, Zhang J, Xiang Y (2020) Data-driven cyber security in perspective-intelligent traffic analysis. IEEE Trans Cyber 50(7):3081–3093. https://doi.org/10.1109/TCYB.2019.2940940
12. Yang H-M, Zhang X-Y, Yin F, Yang Q, Liu C-L (2020) Convolutional prototype network for open set recognition, pp 1–1. https://doi.org/10.1109/tpami.2020.3045079
13. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: CVPR 2011, pp 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347
14. Wang Y, Salamon J, Bryan NJ, Pablo Bello J (2020) Few-shot sound event detection. In: ICASSP 2020:81–85
15. Ntalampiras S (2021) Speech emotion recognition via learning analogies 144:21–26. https://doi.org/10.1016/j.patrec.2021.01.018
16. Laso PM, Brosset D, Puentes J (2017) Dataset of anomalies and malicious acts in a cyber-physical subsystem. Data Brief 14:186–191. https://doi.org/10.1016/j.dib.2017.07.038
17. European commission: white paper on artificial intelligence: a European approach to excellence and trust. Technical report, Brussels (19 February 2020)
18. Alippi C, Ntalampiras S, Roveri M (2016) Online model-free sensor fault identification and dictionary learning in cyber-physical systems. In: 2016 International joint conference on neural networks (IJCNN), pp 756–762. https://doi.org/10.1109/IJCNN.2016.7727276
19. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R (1994) Signature verification using a "siamese" time delay neural network. In: Cowan JD, Tesauro G, Alspector J (eds) Advances in neural information processing systems 6, pp 737–744. Morgan-Kaufmann
20. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP), pp 1–6. https://doi.org/10.1109/MLSP.2015.7324337
21. Purwins H, Li B, Virtanen T, Schlüter J, Chang S, Sainath T (2019) Deep learning for audio signal processing. IEEE J Selected Top Signal Process 13(2):206–219. https://doi.org/10.1109/JSTSP.2019.2908700
22. Koch G, Zemel R, Salakhutdinov R (2015) Siamese neural networks for one-shot image recognition
23. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories 28(4):594–611. https://doi.org/10.1109/tpami.2006.79
24. Taylor P (2006) The target cost formulation in unit selection speech synthesis. In: INTERSPEECH 2006, Pittsburgh, PA, USA. 17–21 Sept 2006
25. Vitanyi PMB (2011) Information distance in multiples 57(4):2451–2456. https://doi.org/10.1109/tit.2011.2110130
26. Feng L, Zhao C (2021) Fault description based attribute transfer for zero-sample industrial fault diagnosis. IEEE Trans Ind Inf 17(3):1852–1862. https://doi.org/10.1109/TII.2020.2988208
27. Acconcjaioco M, Ntalampiras S (2021) One-shot learning for acoustic identification of bird species in non-stationary

environments. In: 2020 ICPR, pp 755–762. https://doi.org/10.1109/ICPR48806.2021.9412005
28. Ntalampiras S (2015) Fault identification in distributed sensor networks based on universal probabilistic modeling. IEEE Trans Neural Netw Learn Syst 26(9):1939–1949. https://doi.org/10.1109/TNNLS.2014.2362015
29. Garbuk SV (2018) Intellimetry as a way to ensure ai trustworthiness. In: 2018 International conference on artificial intelligence applications and innovations (IC-AIAI), pp 27–30
30. Xu G, Li H, Liu S, Yang K, Lin X (2020) Verifynet: Secure and verifiable federated learning. IEEE Trans Inf Forensics Security 15:911–926. https://doi.org/10.1109/TIFS.2019.2929409
31. Capgemini (2021) Reinventing cybersecurity with artificial intelligence: A new frontier in digital security. Technical report, Capgemini Research Institute. https://www.capgemini.com/research/reinventing-cybersecurity-with-artificial-intelligence/
32. Baltrušaitis T, Ahuja C, Morency L-P (2019) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443. https://doi.org/10.1109/TPAMI.2018.2798607