Received: 13 October 2019 | Revised: 19 January 2020 | Accepted: 1 April 2020

DOI: 10.1111/exsy.12564

SPECIAL ISSUE PAPER

Expert Systems WILEY

Real time detection of acoustic anomalies in industrial processes using sequential autoencoders

Barış Bayram | Taha Berkay Duman | Gökhan Ince

Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey

Correspondence
Gökhan Ince, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul, Turkey.
Email: gokhan.ince@itu.edu.tr

Abstract

Development of intelligent systems with the pursuit of detecting abnormal events in real world and in real time is challenging due to difficult environmental conditions, hardware limitations, and computational algorithmic restrictions. As a result, degradation of detection performance in dynamically changing environments is often encountered. However, in the next-generation factories, an anomaly detection system based on acoustic signals is especially required to quickly detect and interfere with the abnormal events during the industrial processes due to the increased cost of complex equipment and facilities. In this study we propose a real time Acoustic Anomaly Detection (AAD) system with the use of sequence-to-sequence Autoencoder (AE) models in the industrial environments. The proposed processing pipeline makes use of the audio features extracted from the streaming audio signal captured by a single-channel microphone. The reconstruction error generated by the AE model is calculated to measure the degree of abnormality of the sound event. The performance of Convolutional Long Short-Term Memory AE (Conv-LSTMAE) is evaluated and compared with sequential Convolutional AE (CAE) using sounds captured from various industrial manufacturing processes. In the experiments conducted with the real time AAD system, it is shown that the Conv-LSTMAE-based AAD demonstrates better detection performance than CAE model-based AAD under different signal-to-noise ratio conditions of sound events such as explosion, fire and glass breaking.

KEYWORDS

acoustic anomaly detection, audio feature extraction, convolutional autoencoder, convolutional long short-term memory autoencoder, industrial processes

1 | INTRODUCTION

The usage of smart systems in homes, factories, cities, and so forth, become more popular to ease the life of humans, especially in surveillance and monitoring tasks. Therefore, a wide variety of sensory information of different type and nature stemming from vision, audition, force/torque, temperature, energy consumption, power, network, and so forth, are individually or jointly utilized in monitoring tasks. However, processing the signals in real time is a challenging problem for abnormal event detection in dynamically changing environments.

The aim of anomaly detection is to distinguish abnormal events from the usual ones. For the new generation of industrial manufacturing systems, monitoring of production with a focus on anomalies is one of the significant capabilities, since abnormal events can affect the quality of manufactured products, deteriorate the continuity and the reliability of the processes and assets (Panfilenko, Poller, Sonntag, Zillner, & Schneider, 2016). Even worse, some anomalies in production processes can endanger the safety of people who use industrial machines in the

factory. Emergency situations and dangers have to be detected and avoided before they become hazardous to ascertain the safety of the workers (Sonntag, Zillner, vander Smagt, & Lörincz, 2017). Therefore, anomaly detection problem is frequently investigated in the framework of smart factories.

The existing anomaly detection systems used in the industrial domain are mainly tailor-made depending on the properties of sensors, sensory information, environmental conditions, and quality metrics of the manufactured products. Among those systems, most common visual anomaly detection systems have some drawbacks such as illumination, occlusion by objects, being out of the field of view, and so forth, which strongly affect the performance of the system (Sodemann, Ross, & Borghetti, 2012). Also, most of the systems processing the videos for monitoring of production require high computation power to achieve real time performance. The Acoustic Anomaly Detection (AAD) systems, however, are not affected by the aforementioned problems; thus offer an advantage using acoustic data features. The audio signals captured by a single-channel microphone, which is available almost on all tablets, mobile phones, augmented reality helmets and headsets as well as on low-end computational sensor systems can be processed in real time with relatively lower computational cost.

To tackle AAD in the industrial domain, there are only few works proposing the use of deep learning techniques. In our previous work proposed an approach based on an unsupervised deep network for AAD to be used in the industrial setting. However, the system was relying on the framewise analysis of the sounds in offline manner instead of representing them in time-series. Also, there is no study using unsupervised sequential learning methods on time-series data in order to detect acoustic anomalies in real time. Besides, the Autoencoders (AEs) are used especially in vision tasks such as abnormal activity recognition in videos, since they are able to determine the spatial and time-frequency information. Thus, in this study, we propose a real time AAD using a Convolutional Long Short-Term Memory AE (Conv-LSTMAE) model (Xingjian et al., 2015), which is a sequence-to-sequence AE, working on feature sequences composed of spectrograms, generated from an audio stream.

Also, considering the real time AAD system, we adapt the Convolutional AE (CAE) to work in a sequential way by using 3D convolutional layers for the 2D features with timestamps to model the feature sequences. The AEs will be evaluated not only in terms of detection performance, but also computation time of the detection to evaluate the compliance with the real time constraints.

2 | RELATED WORK

2.1 | Anomaly detection for industrial purposes

In the literature, several studies have been presented using machine learning techniques for anomaly detection in industrial tools, applications, and processes of manufacturing. The surveys (Bayar, Darmoul, Hajri-Gabouj, & Pierreval, 2015; Diez-Olivan, Del Ser, Galar, & Sierra, 2019) discuss the fault and anomaly detection methods applied in industrial processes. For the anomaly, or novelty detection, One-Class Support Vector Machine (OCSVM) is one of the most widely used detection techniques, and also used in various industrial applications such as kinematic chains (Cariño-Corrales et al., 2016), steering systems (Carino et al., 2016), industrial arms (Narayanan & Bobba, 2018), and so forth. Another study (Jia, Zhao, Di, Yang, & Lee, 2018) presents the comparison of the performances of anomaly detection techniques in industrial applications that are unsupervised Nearest Neighbor, OCSVM, Local Outlier Factor (LOF), Principal Component Analysis, and Maximum Mean Discrepancy.

In Jager and Hamprecht (2009), an approach for monitoring of laser welding processes is developed using Principle Component Analysis and Hidden Markov Model on high dimensional data in order to detect erroneous cases. Also, a kernel-based approach using Support Vector Data Description for novelty detection has been proposed for machinery components used in the industrial domain (Wang, Yu, Lapira, & Lee, 2013). Żabiński, Mączka, and Kluska (2017) applied LOF and neural network (NN) besides OCSVM. In order to detect anomalies in smart factory concept, intelligent solutions based on NNs (Nguyen, Van Ma, & Kim, 2018; Sonntag et al., 2017) are proposed. Furthermore, generalized extreme value distribution is exploited with an anomaly detection algorithm by processing the acoustical signature of the welding process (Hartman, 2012). These works related to industrial AAD only focused on product quality and equipment failure instead of end-to-end industrial processes or applications. In this article, we also propose to employ unsupervised anomaly detection methods on the acoustic data gathered directly from industrial processes and plants. AAD: ACOUSTIC ANOMALY DETECTION

Apart from the traditional machine learning approaches, it has been shown in a few works that deep learning-based methods have improved the performance for the anomaly detection in industrial tasks. The works exploiting deep learning have utilized AEs for abnormal event detection in manufacturing processes. Using the real data of temperature, stream flow, the shocks of machines in the steel factory, Bae, Jang, Kim, and Joe (2018) aim to develop a method based on AE for anomaly detection. However, there are several anomaly detection works based on AEs, used in various environments, but industrial one, explained in the next section. In industrial manufacturing processes, our previous work (Duman, Bayram, & İnce, 2019) presents an approach based on an AE to detect the abnormal events using audio features. Also, it is shown that this approach demonstrates better AAD performance compared to both OCSVM, and a hybrid model of OCSVM and CAE. In this article, we propose a deep network approach based on AE for real time AAD, to be used in industrial manufacturing processes.

2.2 | AEs for anomaly detection

With the advances in deep learning in recent years, AEs are used widely in anomaly detection tasks utilizing acoustic data. A research presented an anomaly detection algorithm by extending Variational AE to apply a supervised method on air conditioner failure detection (Kawachi, Koizumi, & Harada, 2018). For capturing spatio-temporal information, an AE composed of 3D convolutional layers have been utilized in Zhao et al. (2017) to encode the changes of the information to summarize the motion in videos for abnormal activity detection. Therefore, consecutive frames of fixed size as a sequence are stacked by using the 3D kernel that produces the 3D convolutional feature map in which the temporal information constitutes the third dimension.

In another anomaly detection study, the convolutional LSTM has been utilized for surveillance purposes to detect abnormal events in videos, and several competing unsupervised representation learning models including Conv-LSTMAE have been reviewed for anomaly detection on spectrograms and videos (Kiran, Thomas, & Parakkal, 2018). The Conv-LSTM units are evaluated on the anomaly detection datasets by modeling long video sequences (Chong & Tay, 2017). Therefore, an AE based on a composition of Conv-LSTMs with three Conv-LSTM layers is exploited to learn the normal activities in the videos. Moreover, using thermal camera, the unsupervised deep network, Conv-LSTMAE is exploited to detect unseen falls (Nogas, Khan, & Mihailidis, 2018), and the detection is achieved using the mean and standard deviation of reconstruction errors across contexts to estimate the highly variable error. In this study, the AE is compared with CAE and Denoising AE in terms of the metric defined as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). In our work, the sequences generated from spectrograms will be utilized to train Conv-LSTMAE and CAE for AAD on streaming signal in real time.

2.3 | Acoustic anomaly detection

For abnormal event detection using acoustic signals, mostly supervised sequential methods have been utilized like in Kim, Jeon, and Kim (2019) and Hayashi, Komatsu, Kondo, Toda, and Takeda (2018). However, only few approaches for AAD have been proposed for different application domains such as surveillance using a method based on OCSVM with a Radial Basis Function kernel (Aurino et al., 2014). In these works (Marchi, Vesperini, Squartini, & Schuller, 2017; Principi, Vesperini, Squartini, & Piazza, 2017), different types of AEs are proposed for acoustic novelty detection, including regular AE, Compression AE, Denoising AE, and Adversarial AE. Also, the research in Koizumi, Saito, Uematsu, Kawachi, and Harada (2019) proposed an unsupervised AAD system using AE by applying an optimization principle to maximize True Positive Rate (TPR) while minimizing False Positive Rate (FPR). Droghini et al. presented a CAE-based AAD algorithm through the end-to-end strategy for human fall detection in an indoor environment (Droghini, Ferretti, Principi, Squartini, & Piazza, 2017). They compared the performance of the proposed system with OCSVM and showed a significant improvement using CAE. However, in the studies, the anomaly detection was not applied for industrial purposes. Therefore, in our previous work (Duman et al., 2019), we presented the use of AE for AAD in industrial tasks, and compared with OCSVM using original audio datasets. In that work, it was shown that a deep network, CAE provides better AAD results than OCSVM, but it was an offline framewise solution not capable of working in real time. Thus, in this article, we focus on the sequential AEs to process the acoustic signals in real time using time-series representation.

La previsione dell'ordine è abbastanza importante nell'ambito di anomaly detection in ambiente industriale.

3 | AAD SYSTEM

The proposed AAD system shown in Figure 1 consists of two main phases: (1) An offline training phase responsible of the generation of sequential AE model and (2) an online operation phase to conduct the AAD in real time. In the feature extraction module, sliding window operation on the streaming signal is applied, and the Mel-spectrograms are created, which is a common module for both the training and the AAD phases. The sliding window operation is applied on the prerecorded audio files in the training phase, whereas using the same step size and window length it is applied on the incoming audio signals captured by the microphone in real time. In the model generation module of the training phase, feature sequences are used to generate a model from stored audio files. In the AAD module in the operation phase, however, the reconstruction of the sequence is conducted followed by the reconstruction error calculation. Using a final thresholding operation, the event is discriminated as normal or anomaly event.

3.1 | Feature extraction

In the proposed AAD system, feature extraction is a common module to be used for the extraction of audio features and generation of feature sequences. Individual components of this module, a sliding window operation, Mel-spectrogram extraction and sequence generation, are utilized

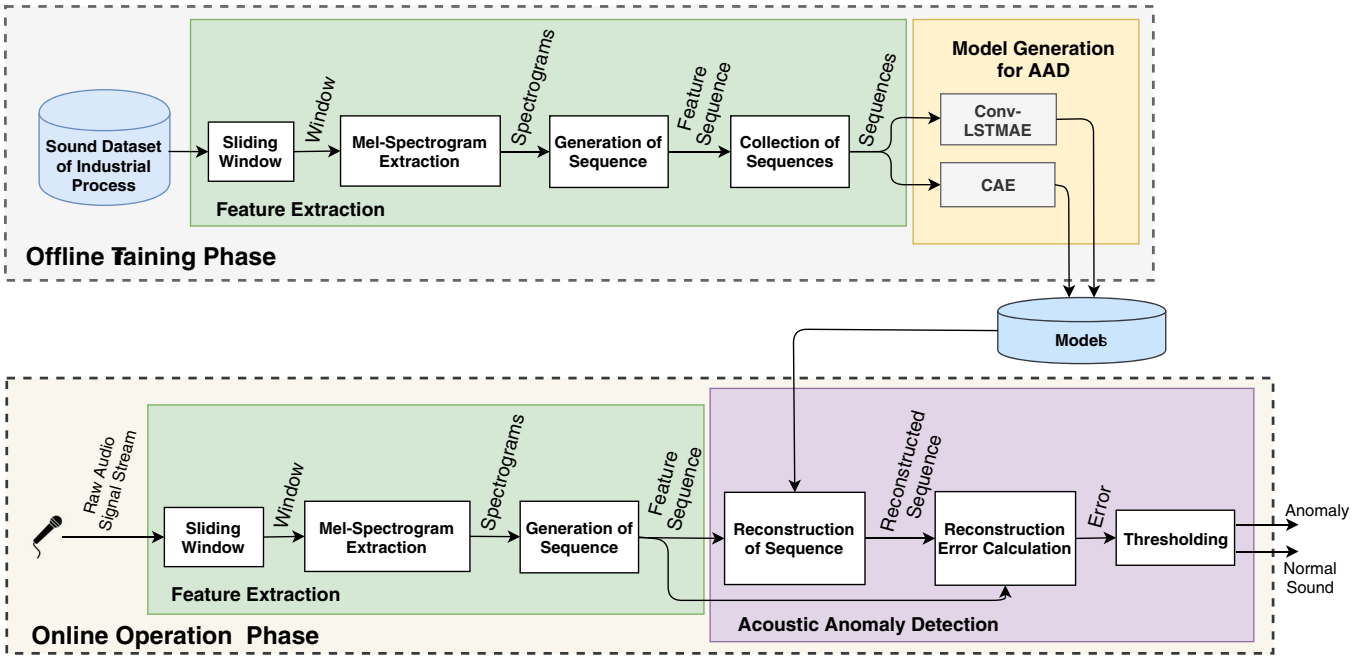

FIGURE 1 Overview of the acoustic anomaly detection system

in both of the phases. In the offline training phase, an additional component for the collection of sequences exists to create the training and validation sets for the module regarding the model generation of AAD.

The sliding window operation is applied to the raw signals of the recorded audio files in the offline training phase, or to the audio stream from a microphone in the online operation phase in order to generate windows with a step size, S, and a window length, L all applied in time domain.

Thus, a window overlaps with a length, $L - S$, with the previous window. After composing a window, a Mel-scaled spectrogram, a 2D time-frequency representation of sound, is extracted from it as a feature. Then, the features of a given number (N) consecutive windows are combined to generate a sequence. A sequence includes the new window at time t having all windows between the windows at time $t - (L - 1)$ and t. Repeating the same processing chain for all audio signals, the sequences generated from audio recordings stemming from industrial processes are collected and used for the training of the AE models.

On the other hand, in the online operation phase, the windows are acquired from the streaming signal captured by a microphone. When a new incoming window is obtained after the sliding window procedure, the sequence is updated in which the first window is removed, and the new one is inserted to the end of this sequence so that a new sequence including the new window is generated.

3.2 | Model generation for AAD

An AE is used for encoding input into fewer dimensions and afterwards it reconstructs the same input from the encoded representation (Bengio, Courville, & Vincent, 2013; Hinton & Salakhutdinov, 2006). It resembles a NN consisting of two components, encoder and decoder. Encoder is used for encoding the input and dimensional reduction, whereas the decoder aims to regenerate the same input as accurately as possible. For a sequential AE, the input and output of the encoder and decoder are sequences having a fix-length and being composed of audio features.

In the model generation module, a sequential AE is trained using the natural and known sounds of the industrial processes. Major part of the generated sequences is utilized for training and validation of AE model, and the rest of the sequences are exploited for testing in order to evaluate the model and to estimate the threshold for AAD module.

In this study, we implement two unsupervised Deep NNs (Figure 2) used for sequence-to-sequence encoding and decoding operations, and investigate the effectiveness of the conventional approach of CAE used in sequential manner and Conv-LSTMAE that is based on Recurrent NN. The inputs of the networks are spectrograms with 128×128 size, and each sequence consists of five spectrograms. The purple and blue-colored layers are encoding and decoding parts, respectively. The 3D version of max pooling and up sampling layers of CAE network are utilized to reduce and improve the shape of the data, respectively. Also, Conv-LSTMAE has three additional LSTM layers colored by yellow in which first two layers belong to the encoding step, and the third one belongs to the decoding step.

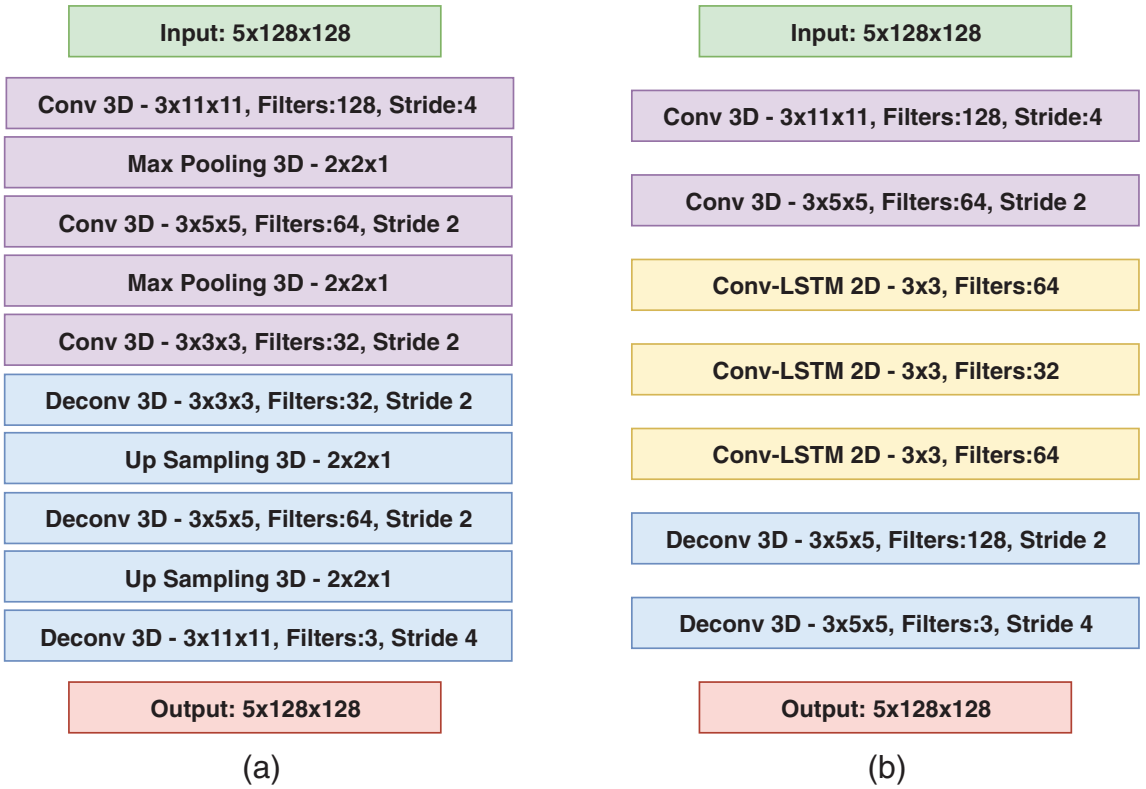

FIGURE 2 The sequential AEs used for real time AAD: (a) the architecture of CAE based on convolutional layers, max pooling, and up sampling layers and (b) the architecture of Conv-LSTMAE composed of 2D LSTM layers based on recurrent neural network and convolutional layers

3.2.1 | Convolutional AE

One of the AEs applied on the spectrograms is CAE in a sequential manner. In several studies, it is stated that CAE provides better performances in learning of multivariate data for extraction of embedding representations, anomaly detection, novelty detection, image compression, image denoising, and so forth. CAE is fundamentally based on consisting of multiple convolutional layers like in Figure 2a instead of fully connected layers to hierarchically extract the features. Also, the encoder network is composed of convolutional layers, and the decoder network has the transposed convolutional layers. For the CAE network, total number of parameters and number of floating-point operations are 1.4 M and 2.8 M, respectively.

It is noteworthy to emphasize the importance of heuristically setting the appropriate hyperparameters such as the number of convolution, pooling, dropout and upsampling layers, the number of filters, the parameters of the optimization algorithm, and the types of the loss function and evaluation metric for CAE. Since we need to reconstruct the spectrograms for the anomaly detection, the Rectified Linear Units (ReLU) is employed as the activation function.

3.2.2 | Convolutional LSTM AE

For modeling of sequences for various tasks such as motion estimation, activity recognition, text generation, and so on, LSTM, a special type of Recurrent NN is utilized. The LSTM cells are able to model varying-length time-series, and to learn the long-term dependencies. For unsupervised deep learning, LSTM-based AEs have been utilized for sequential feature extraction, and abnormal event detection. Instead of matrix multiplication, convolutional filters are utilized for Conv-LSTM network, thus Conv-LSTM basically is a composite LSTM that can be utilized for two-dimensional sequences like videos or audio signals. A streaming audio signal has both spectral and temporal information, therefore we may utilize Conv-LSTM to learn the signals considering both of information. Also, for the training of Conv-LSTMAE model ReLU activation function is utilized, and the appropriate number of LSTM layers, filter and kernel sizes are heuristically set as in Figure 2b. Total number of parameters and number of floating-point operations of this LSTM network are 1.4 M and 2 M, respectively.

3.3 | Acoustic anomaly detection

3.3.1 | Reconstruction of feature sequences

In the reconstruction stage as illustrated in Figure 3, new sequence of spectrograms is reconstructed in encoding-decoding layers of the model. The AE model takes the new sequence as input and reconstructs it to a sequence with the same size as output. The CAE and Conv-LSTMAE both take a spectrogram feature from the sequence, and then reconstruct it at a time. After L time, the sequence will be reconstructed. The aim of an AE model is to generate a reconstructed sequence so close to the actual one, so the similarity distance between them gives the abnormality degree of the sequence.

3.3.2 | Calculation of reconstruction error

To estimate the reconstruction error, Euclidean distance is calculated for each actual spectrogram and the reconstructed spectrogram in the corresponding sequences. The average of the distance between the sequences gives the reconstruction error (Principi et al., 2017) as in Equation (1).

$$E_R = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(X_i^a - X_i^r\right)^T\left(X_i^a - X_i^r\right), \quad (1)$$

where E_R is the reconstruction error, N_s denotes the number of spectrogram features in the sequence, X^a is the actual spectrogram, X^r is the reconstructed spectrogram, ^{T} is the transpose operator, and i represents the index of the spectrogram feature belonging to the sequence. The reconstruction error is checked against the threshold obtained in the testing step of the model generation stage using Youden's index, J, as in

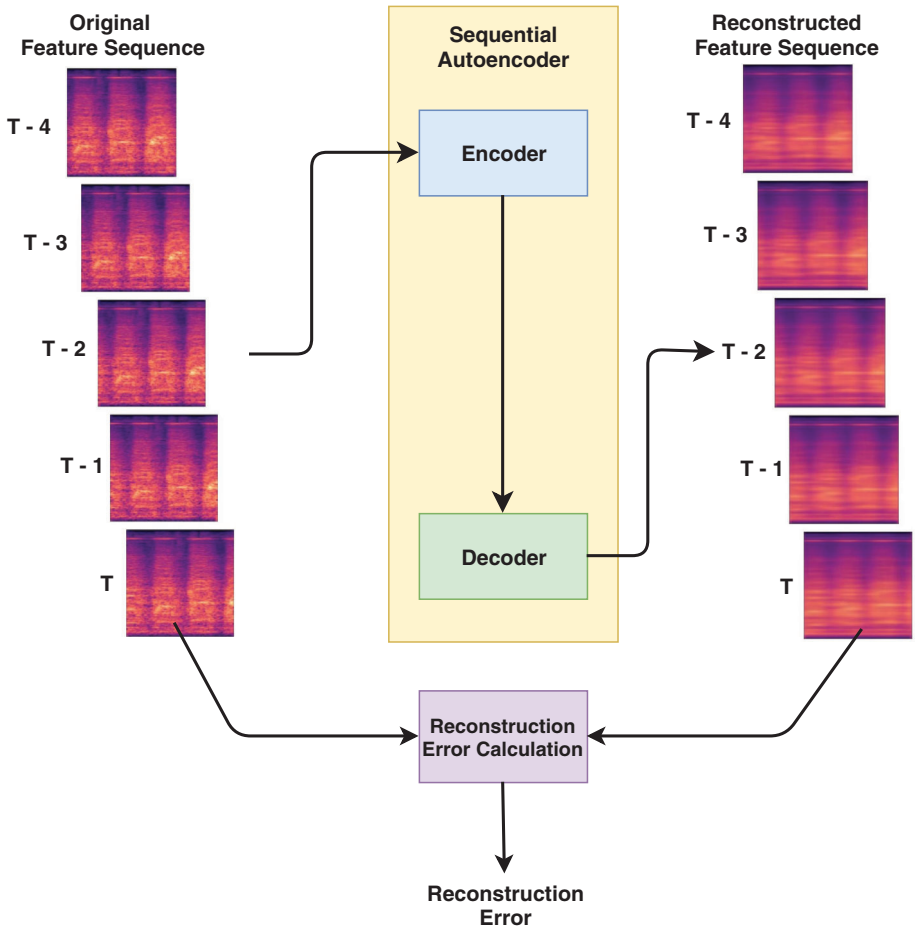

FIGURE 3 The reconstruction and reconstruction error calculation stages

Equation (2). This value is applied on each threshold present in the ROC-AUC (Receiver Operating Characteristic) curve estimated using the Euclidean distance values (Youden, 1950) in order to estimate the best threshold with the maximum J value as follows;

$$J = \text{Sensitivity} + \text{Specificity} - 1, \tag{2}$$

where *Sensitivity* stands for TPR and *Specificity* stands for True Negative Rate. The aim of Youden's index is to maximize the difference between TPR and FPR. Thus, the system will set the optimal threshold value by maximizing the anomaly detection accuracy and separability rate.

4 | EXPERIMENTS AND RESULTS

4.1 | Experimental setup

We created a new acoustic dataset for the industrial processes and plants by applying the two main steps of dataset creation process (Fonseca et al., 2019):

- Data collection: We retrieved audio data from Youtube videos recorded during industrial processes as well as anomaly sounds.
- Data arrangement: We removed the empty (silent), noisy (background noise being too high and interfering too much with the target sound quality), and irrelevant parts from the original audio data.

It is hard to capture abnormal patterns because of their rarity in real life. In addition to that, creating anomaly events and recording respective sounds are costly. Thus, the sounds of the these processes, including the abnormal patterns in which the normal patterns are mixed with the abnormal ones by fixed signal-to-noise ratio (SNR) values are streamed in an online manner in the testing phase. As a consequence, for generating anomaly patterns, we mixed a number of normal sounds with different types of noises at specific SNRs of {5, 0, −5, −10, −15} dB. It is noteworthy to mention that the useful signals consist of the industrial sounds used in training and the noises are anomaly sounds; therefore the SNR tends to get lower while the power of the anomaly event gets higher and vice versa.

We created four acoustic datasets

as listed in Table 1, which contain sounds involving industrial (i) painting, (ii) cutting, (iii) welding, and (iv) robotic arm sounds. All signals in the dataset are sampled at 22,050 Hz. Using the audio files from these datasets, we selected three anomaly sounds to be mixed for generating anomaly events; *Explosion*, *Fire*, and *Glass Breaking*.

In the offline training and online operation phase, the step size and length of the window are selected as S = 0.05 second and L = 1 second, respectively. Also, the number of windows is set to N = 5 in a sequence. Each AE is trained using the following parameters: 100 – 200 epochs for Conv-LSTMAE and 200 – 500 epochs for CAE using an early stopping criteria, a batch size of 10 as the subset size of the training examples in one forward/backward pass, and Adam optimizer with learning rate of 0.001 as the step size at each epoch to optimize each parameter.

4.1.1 | Hardware and software specifications

The computations for the model generations and the experiments were run on a machine with Intel® Core™ i7-8700K CPU and GeForce GTX 1080Ti GPU. For the extraction of the Mel-spectrogram, Librosa library is utilized. In addition, Keras library is employed for the implementation and application of the CAE network and carrying out the experiments. Also, for real time experiments, HARK (HRI-JP Audition for Robots with Kyoto University) is utilized which is an open source robot audition software.

TABLE 1 Acoustic datasets

Datasets	Total duration in minutes	Number of recordings
Industrial painting	26	16
Industrial cutting	18	15
Industrial welding	28	14
Industrial robotic arm	10	17

4.2 | Evaluation criteria

We select the Area Under the ROC-AUC curve (ROC-AUC) score as the first evaluation metric to determine the degree of separability. In all possible threshold values, it measures the classification performance of the model. With this metric, we evaluate how the model works while distinguishing between normal and anomaly events. The higher the ROC-AUC score is, the better representative the model is. A ROC-AUC score close to 1.0 indicates that the model is perfectly able to separate anomaly and normal events.

We also use F1-score as the second evaluation criterion. It balances precision and recall by giving equal weights to both of them. It indicates the mean of precision and recall for a specific threshold value. This threshold value is specified by Youden's index as in Equation (2) in ROC-AUC. Using F1-score at this threshold, we evaluate the decision making performance of the model.

4.3 | Results

4.3.1 | Performance of AAD

Tables 2-5 show the AAD results of both Conv-LSTMAE and CAE models in terms of mean ROC-AUC and mean F1 scores on four different industrial sound sets including three types of anomalies. The main tendency is that the ROC-AUC and F1-scores drop significantly during industrial events of painting (Table 2) and welding processes (Table 3). Also, the worst ROC-AUC and F1-scores of both of AEs are obtained while monitoring the painting processes.

It is observed that the CAE model has equal or higher F1-scores on the detection of anomalies in the welding processes (Table 3) when any kind of anomaly sound is present. Moreover, CAE provides better ROC-AUC scores for detection of anomalies only during the sounds of cutting processes (Table 4) with the explosion anomaly sound at low SNR levels.

TABLE 2 Anomaly detection results on industrial painting sound dataset

Anomaly	Explosion				Fire				Glass breaking			
Method	Conv-LSTMAE		CAE		Conv-LSTMAE		CAE		Conv-LSTMAE		CAE	
Metric												
SNR (dB)	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score
5	0.51	0.48	0.42	0.43	0.47	0.67	0.45	0.43	0.59	0.47	0.61	0.43
0	0.53	0.50	0.50	0.43	0.51	0.46	0.51	0.44	0.67	0.478	0.68	0.44
-5	0.59	0.51	0.56	0.44	0.52	0.467	0.59	0.44	0.72	0.50	0.70	0.44
−10	0.64	0.53	0.63	0.45	0.64	0.49	0.66	0.45	0.79	0.56	0.74	0.56
−15	*0.78*	0.55	*0.71*	*0.47*	*0.79*	*0.497*	*0.71*	*0.458*	0.86	0.638	*0.76*	0.59

TABLE 3 Anomaly detection results on industrial welding sound dataset

Anomaly	Explosion				Fire				Glass breaking			
Method	Conv-LSTMAE		CAE		Conv-LSTMAE		CAE		Conv-LSTMAE		CAE	
Metric												
SNR (dB)	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score
5	0.54	0.50	0.63	0.47	0.50	0.48	0.51	0.469	0.41	0.467	0.44	0.468
0	0.57	0.60	0.74	0.60	0.54	0.60	0.74	0.58	0.41	0.48	0.51	0.47
−5	0.71	0.61	0.82	0.615	0.67	0.62	0.82	0.60	0.47	0.57	0.74	0.57
−10	0.84	0.625	0.84	0.626	0.88	0.628	0.83	0.618	0.67	0.617	0.81	0.618
−15	0.89	0.639	0.84	*0.64*	0.89	0.64	0.84	0.64	0.86	0.62	0.83	*0.626*

TABLE 4 Anomaly detection results on industrial cutting sound dataset

Anomaly	Explosion				Fire				Glass breaking			
Method	Conv-LSTMAE		CAE		Conv-LSTMAE		CAE		Conv-LSTMAE		CAE	
Metric												
SNR (dB)	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score
5	0.67	0.48	0.66	0.44	0.50	0.455	0.48	0.447	0.55	0.547	0.54	0.47
0	0.68	0.52	0.66	0.447	0.62	0.457	0.52	0.45	0.55	0.55	0.54	0.51
−5	0.68	0.52	*0.71*	0.448	0.63	0.47	0.53	0.45	0.62	0.57	0.54	0.52
−10	0.79	0.56	*0.84*	0.54	0.72	0.51	0.55	0.455	0.64	0.63	0.60	0.52
−15	0.87	0.59	*0.90*	0.557	0.85	0.57	0.61	0.555	0.76	0.69	0.71	0.58

TABLE 5 Anomaly detection results on robotic arm movement sound dataset

Anomaly	Explosion				Fire				Glass breaking			
Method	Conv-LSTMAE		CAE		Conv-LSTMAE		CAE		Conv-LSTMAE		CAE	
Metric												
SNR (dB)	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score	ROC AUC	F1 score
5	0.72	0.63	0.41	0.46	0.72	0.46	0.46	0.495	*0.92*	0.72	0.78	0.717
0	*0.82*	*0.807*	0.58	0.51	*0.86*	0.48	0.66	0.547	0.95	0.807	0.91	0.88
−5	0.94	0.93	0.74	0.709	0.96	0.51	0.94	0.795	0.97	0.93	0.97	0.91
−10	0.97	0.98	0.97	0.947	0.97	0.737	0.97	0.91	0.98	0.98	0.99	0.92
−15	*0.98*	*1.0*	*0.98*	*0.99*	*0.98*	*0.94*	*0.99*	*0.951*	*0.98*	*1.0*	*0.99*	*0.988*

At high levels of SNR, the only acceptable performance (Table 5) is obtained by Conv-LSTMAE on the detection of anomalies while the sounds of the industrial robotic arm movement are present. At SNR = 0 dB, the explosion anomalies are easily detected indicated by the mean ROC-AUC and mean F1-scores being 82% and 80.7%. Also, the mean ROC-AUC scores are 72% for SNR = 5 dB and 86% for SNR = 0 dB, respectively, when the sounds of explosion and fire are superimposed as anomaly. Moreover, the highest mean ROC-AUC and mean F1-scores are estimated during the robotic arm motions at low SNR levels.

In general, Conv-LSTMAE outperforms the CAE model and it detects the anomalies with higher mean ROC-AUC and mean F1-scores even in the low SNRs such as 0 dB and −5 dB. However, when the SNR is 5 dB, the natural industrial process sounds are mostly dominated by noises, therefore both AEs yield similar performances. It is observed that the Conv-LSTMAE has higher F1-scores than CAE; but at high SNR levels, such as SNR = 0 dB and SNR = 5 dB, most of the scores are rather low for both AE. Moreover, the average of the computational times including sliding window operations, extraction of five spectrograms, generation of one sequence, and reconstruction of the sequence for Conv-LSTMAE and CAE are 26 milliseconds and 23 milliseconds in average, respectively. The times are appropriate for real time operation of AAD using both AE models.

4.3.2 | Discussion

Overall, at all levels of SNR, Conv-LSTMAE achieves the best scores for the fire and glass breaking anomalies in the painting, welding, and cutting processes. Also, on the sounds robotic arm movements, the anomalies are robustly detected by Conv-LSTMAE. However, at the level of SNR being 5 dB, 0 dB, and −5 dB, the performances demonstrated by the AAD for the explosion sounds during welding and painting processes are not impressive, since the sounds of those processes are vaguely apparent and utterly close to white noise. However, at lower SNR values, the anomalies are easily detected. Also, the power spreads of glass breaking sound across all frequencies within the spectrogram in contrast to the other anomaly sounds. Thus, it can be seen that the glass breaking is the most distinguishable anomaly sound by both of the AE methods. The anomaly is easily detected at each level of SNR with high ROC-AUC score.

On the other hand, the power of the explosion and fire spectral patterns are concentrated in the lower frequencies, and as the frequency increases, it can be said that the spread pattern gets similar to white noise, and at high frequencies this effect vanishes. For these reasons, the glass breaking anomaly can be detected with higher accuracy than other anomalies.

In addition, based on the analysis carried out on the TPRs and FPRs obtained in all experiments, the ROC-AUC curves at each level of SNR are given in Figure 4 for Conv-LSTMAE on the left and CAE on the right. To reiterate, Conv-LSTMAE has better mean ROC-AUC scores, and Conv-LSTMAE obviously performs better than the CAE at each level of SNR because its Pareto efficiency is higher.

Because the sounds have recurrent time-series patterns, the Conv-LSTMAE model can robustly model them as shown in the left panel of Figure 5, so it can discriminate normal and abnormal patterns more distinctly compared to the CAE in the right panel of Figure 5. The errors regarding the sequences including the anomaly sounds (from the sequence number 176–230) are higher in Conv-LSTMAE model than the errors calculated by reconstruction output of CAE model in the same sequence interval, which contributes to improved performance. In the graph regarding Conv-LSTMAE model, the reconstruction errors of the normal robotic arm sound patterns are steady (from the sequence number 0 to 175), whereas the CAE model that provides close ROC-AUC scores in the other three industrial sounds, has difficulty in coping with robotic arm sounds, thus large errors are calculated (between the sequence number 40 and 110, and the sequence numbers larger than 250). As a consequence, in the

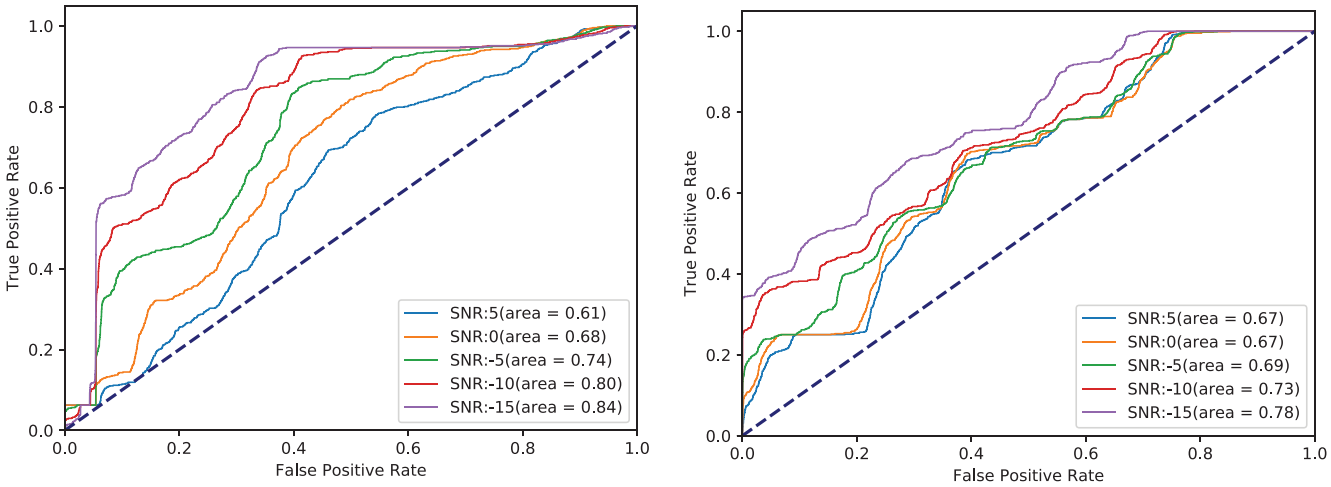

FIGURE 4 Average ROC-AUC curves (TPR vs. FPR) of AAD performances of Conv-LSTMAE (on the left) and CAE (on the right)

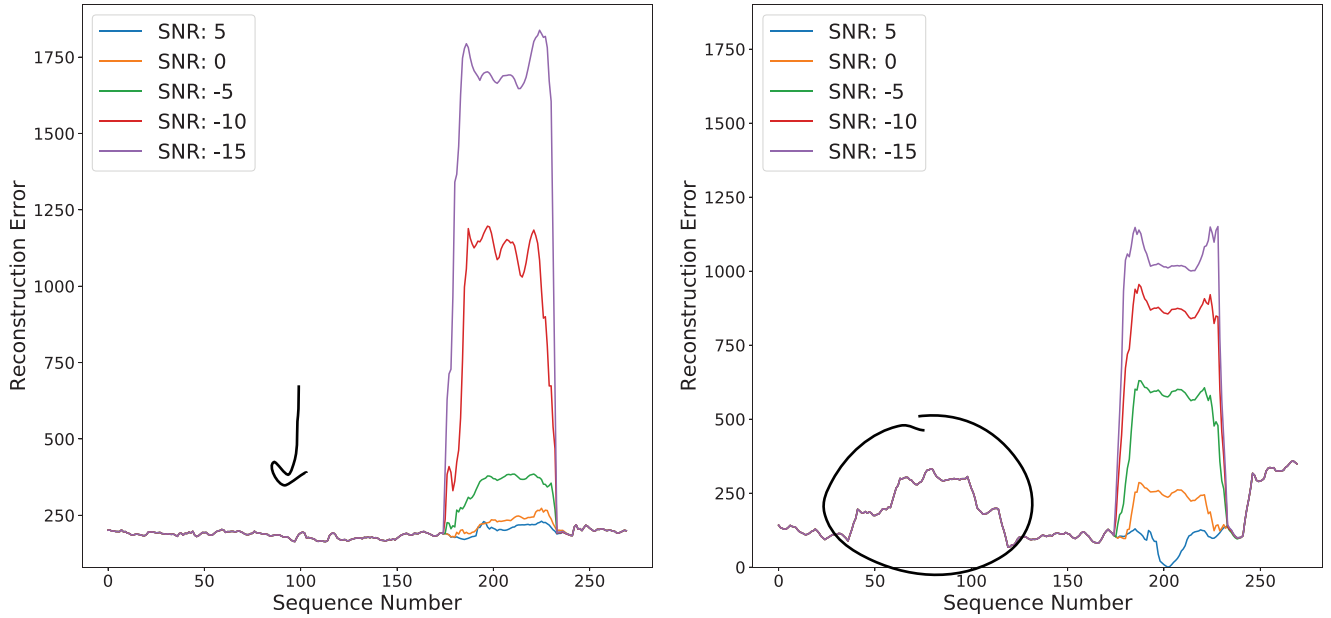

FIGURE 5 The reconstruction results of Conv-LSTMAE (left) and CAE (right) on the signal of industrial robotic arm sound with an explosion anomaly sound

AAD results of the explosion sound in Table 5, the performance of the Conv-LSTMAE model is better than the CAE model at all SNR levels. It is also noteworthy to mention that all anomaly sounds used in the experiments are non-sequential tones, because natural anomalies occurring alongside industrial sounds do not consist of sequential tones, sequences of single frequencies. However, LSTMs are already known to be performing well in the discrimination of stationary sounds having abrupt changes of spectral power in discrete time intervals comprising sequence tones (Yue, Fu, & Liang, 2018).

5 | CONCLUSION

In this study, an approach for the real time AAD using Convolutional Conv-LSTMAE was proposed. This sequential AE was benchmarked against with a CAE model used in sequential manner in terms of performances of anomaly detection and computational time. The performance of the presented detection approach and its ability to distinguish normal sounds and anomalies at different levels of SNR have been evaluated in various industrial datasets created within the framework of this study. We have shown the effectiveness of the Conv-LSTMAE-based AAD running in real time especially in the datasets including industrial arm movement and cutting processes.

In the future, the dataset will be expanded by collecting more and diverse normal and anomaly event sounds from industrial processes. Also, online learning will be achieved by updating the sequential AE model after each reconstruction.

ORCID

Gökhan Ince https://orcid.org/0000-0002-0034-030X

ENDNOTES

[1] https://kovan.itu.edu.tr/index.php/s/MN2jtWCovolkeGi/download.

[2] https://librosa.github.io/librosa.

[3] https://www2.hark.jp/.

REFERENCES

Aurino, F., Folla, M., Gargiulo, F., Moscato, V., Picariello, A., Sansone, C. 2014. *One-class SVM based approach for detecting anomalous audio events*. Paper presented at Proceedings of the 2014 International Conference on Intelligent Networking and Collaborative Systems (pp. 145–151). IEEE.

Bae, G., Jang, S., Kim, M., Joe, I. (2018). *Autoencoder-based on anomaly detection with intrusion scoring for smart factory environments*. Paper presented at Proceedings of the International Conference on Parallel and Distributed Computing: Applications and Technologies (pp. 414–423). Springer.

Bayar, N., Darmoul, S., Hajri-Gabouj, S., & Pierreval, H. (2015). Fault detection, diagnosis and recovery using artificial immune systems: A review. *Engineering Applications of Artificial Intelligence*, *46*, 43–57.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Carino, J. A., Delgado-Prieto, M., Zurita, D., Millan, M., Redondo, J. A. O., & Romero-Troncoso, R. (2016). Enhanced industrial machinery condition monitoring methodology based on novelty detection and multi-modal analysis. *IEEE Access*, *4*, 7594–7604.

Cariño-Corrales, J. A., Saucedo-Dorantes, J. J., Zurita-Millán, D., Delgado-Prieto, M., Ortega-Redondo, J. A., Alfredo Osornio-Rios, R., & de Jesus Romero-Troncoso, R. (2016). Vibration-based adaptive novelty detection method for monitoring faults in a kinematic chain. *Shock and Vibration*, *2016*, 1–12.

Chong, Y.S., Tay, Y.H. 2017. *Abnormal event detection in videos using spatio-temporal autoencoder*. Paper presented at International Symposium on Neural Networks (pp. 189–196). Springer.

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. *Information Fusion*, *50*, 92–111.

Droghini, D., Ferretti, D., Principi, E., Squartini, S., Piazza, F. 2017. *An end-to-end unsupervised approach employing convolutional neural network autoencoders for human fall detection*. Paper presented at Proceedings of the Italian Workshop on Neural Nets (pp. 185–196). Springer.

Duman, T.B., Bayram, B., İnce, G. 2019. *Acoustic anomaly detection using convolutional autoencoders in industrial processes*. Paper presented at Proceedings of the International Workshop on Soft Computing Models in Industrial and Environmental Applications (pp. 432–442). Springer.

Fonseca, E., Plakal, M., Ellis, D.P., Font, F., Favory, X., Serra, X. (2019). Learning sound event classifiers from web audio with noisy labels. arXiv preprint arXiv:1901.01189.

Hartman, D.A. (2012). *Real-time detection of processing flaws during inertia friction welding of critical components*. Paper presented at ASME Turbo Expo 2012: Turbine Technical Conference and Exposition (pp. 1–10). American Society of Mechanical Engineers.

Hayashi, T., Komatsu, T., Kondo, R., Toda, T., Takeda, K. 2018. *Anomalous sound event detection based on wavenet*. Paper presented at Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2494–2498). IEEE.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Jager, M., & Hamprecht, F. A. (2009). Principal component imagery for the quality monitoring of dynamic laser welding processes. *IEEE Transactions on Industrial Electronics*, *56*(4), 1307–1313.

Jia, X., Zhao, M., Di, Y., Yang, Q., & Lee, J. (2018). Assessment of data suitability for machine prognosis using maximum mean discrepancy. *IEEE Transactions on Industrial Electronics*, *65*(7), 5872–5881.

Kawachi, Y., Koizumi, Y., Harada, N. 2018. *Complementary set variational autoencoder for supervised anomaly detection*. Paper presented at Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2366–2370). IEEE.

Kim, N. K., Jeon, K. M., & Kim, H. K. (2019). Convolutional recurrent neural network-based event detection in tunnels using multiple microphones. *Sensors*, *19*(12), 2695.

Kiran, B., Thomas, D., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, *4*(2), 36.

Koizumi, Y., Saito, S., Uematsu, H., Kawachi, Y., & Harada, N. (2019). Unsupervised detection of anomalous sound based on deep learning and the Neyman–Pearson lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(1), 212–224.

Marchi, E., Vesperini, F., Squartini, S., & Schuller, B. (2017). Deep recurrent neural network-based autoencoders for acoustic novelty detection. *Computational Intelligence and Neuroscience*, *2017*, 1–14.

Narayanan, V., Bobba, R.B. 2018. *Learning based anomaly detection for industrial arm applications*. Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy (pp. 13–23). ACM.

Nguyen, V. Q., Van Ma, L., & Kim, J. (2018). Lstm-based anomaly detection on big data for smart factory monitoring. *Journal of Digital Contents Society*, *19*(4), 789–799.

Nogas, J., Khan, S.S., Mihailidis, A. (2018). Deepfall–non-invasive fall detection with deep spatio-temporal convolutional autoencoders. arXiv preprint arXiv: 1809.00977.

Panfilenko, D., Poller, P., Sonntag, D., Zillner, S., Schneider, M. (2016). *Bpmn for knowledge acquisition and anomaly handling in cps for smart factories*. Paper presented at the 201 6 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 1–4). IEEE.

Principi, E., Vesperini, F., Squartini, S., Piazza, F. 2017. *Acoustic novelty detection with adversarial autoencoders*. Paper presented at Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 3324–3330). IEEE.

Sodemann, A. A., Ross, M. P., & Borghetti, B. J. (2012). A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1257–1272.

Sonntag, D., Zillner, S., vander Smagt, P., & Lörincz, A. (2017). Overview of the cps for smart factories project: Deep learning, knowledge acquisition, anomaly detection and intelligent user interfaces. In *Industrial internet of things* (pp. 487–504). New York, NY: Springer.

Wang, S., Yu, J., Lapira, E., & Lee, J. (2013). A modified support vector data description based novelty detection approach for machinery components. *Applied Soft Computing*, *13*(2), 1193–1205.

Xingjian, S., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802–810). Curran Associates, Inc.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Yue, B., Fu, J., & Liang, J. (2018). Residual recurrent neural networks for learning sequential representations. *Information*, *9*(3), 56.

Żabiński, T., Mączka, T., Kluska, J.. (2017). *Industrial platform for rapid prototyping of intelligent diagnostic systems*. Paper presented at Polish Control Conference (pp. 712–721). Springer.

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S. 2017. *Spatio-temporal autoencoder for video anomaly detection*. Paper presented at Proceedings of the 25th ACM international conference on Multimedia (pp. 1933–1941). ACM.

AUTHOR BIOGRAPHIES

Barış Bayram received his B.Sc. degree in Computer Engineering from Izmir University of Economics, Turkey, in 2012 and the M.Sc. degree in Computer Engineering in 2015 from Istanbul Technical University (ITU), Istanbul. He is a Ph.D. candidate at ITU Department of Computer Engineering since 2015. Also, he is currently working in Yapi Kredi Technology as an R&D Engineer on fraud detection. His research interests are signal processing, artificial intelligence, machine learning, deep learning, robotics, and human-robot interaction.

Taha Berkay Duman received his BSc and MSc degrees in Computer Engineering from Istanbul Technical University, Turkey, in 2015 and 2019 respectively. He is currently working on automation processes. His research interests include audio processing, anomaly detection, human robot interaction, industrial automation, and robotics.

Gökhan Ince received the B.S. degree in Electrical Engineering from Istanbul Technical University, Turkey, in 2004, the M.S. degree in Information Engineering in 2007 from Darmstadt University of Technology, Germany and the Ph.D. degree in the Department of Mechanical and Environmental Informatics, Tokyo Institute of Technology, Japan in 2011. From 2006 to 2008, he was a researcher with Honda Research Institute Europe, Offenbach, Germany and from 2008 to 2012, he was with Honda Research Institute Japan, Co., Ltd., Saitama, Japan. Since 2012, he has been an Assistant Professor with the Computer Engineering Department, Istanbul Technical University. His current research interests include humancomputer interaction, robotics, artificial intelligence and signal processing. He is a member of IEEE, RAS, ISAI and ISCA.

How to cite this article: Bayram B, Duman TB, Ince G. Real time detection of acoustic anomalies in industrial processes using sequential autoencoders. *Expert Systems*. 2021;38:e12564. https://doi.org/10.1111/exsy.12564