A deep learning approach for anomaly detection based on SAE and LSTM in mechanical equipment

Zhe Li[1], Jingyue Li [1], Yi Wang[2], Kesheng Wang[3,4]

Abstract Anomaly in mechanical systems may cause equipment to break down with serious safety, environment, and economic impact. Since many mechanical equipment usually operates under tough working environments, which makes them vulnerable to types of faults, anomaly detection for mechanical equipment usually requires considerable domain knowledge. However, a common dilemma in many practical applications is that one may not be able to obtain the empirical knowledge about anomaly or the history data is completely unlabelled, which makes conventional fault identification methods not applicable. In order to fill the gap, this paper proposes a novel deep learning-based method for anomaly detection in mechanical equipment by combining two types of deep learning architectures, stacked autoencoders (SAE) and long short term memory (LSTM) neural networks, to identify anomaly condition in a completely unsupervised manner. The proposed method focuses on the anomaly detection through multiple features sequence when the history data is unlabelled and the empirical knowledge about anomaly is absent. An experiment for anomaly detection in rotary machinery through wavelet packet decomposition (WPD) and data-driven models demonstrates the efficiency and stability of the proposed approach. The method can be divided into two stages: SAE-based multiple features sequence representation and LSTM-based anomaly identification. During the experiment, fivefold cross-validation has been applied to validate the performance and stability of the proposed approach. The results show that the proposed approach could detect anomaly working condition with 99% accuracy under a completely unsupervised learning environment and offer an alternative method to leverage and integrate features for anomaly detection without empirical knowledge.

✉ Zhe Li

zhel@ntnu.no

[1] Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[2] School of Business, Plymouth University, Plymouth Devon, UK

[3] School of Mechanical Engineering, Changzhou University, China

[4] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Keywords Anomaly detection · Mechanical equipment · SAE · LSTM

1. Introduction

Mechanical equipment covers a very broad range of industrial equipment and plays a momentous role in manufacturing application. As the key equipment in many production fields, mechanical machinery usually operates under a tough working environment, which makes it vulnerable to types of faults. These faults may cause equipment to break down or degrade certain machinery performance like geriatric location, manufacturing quality and operation safety [1]. Practically, combined faults would increase the complexity and difficulty of fault classification [2], in which case, companies usually prefer to identify anomaly condition first, and then further diagnosis or prognosis could be conducted.

Considering the complexity of the current industrial applications, anomaly detection in mechanical machinery is a challenging issue nowadays [3,4]. Studies have shown that the human operator is responsible for 70–90% of the accidents in industrial environments [5]. For this reason, computer-based anomaly detection systems with high complexity are imperative to improve the accuracy and reliability of anomaly detection, and prevent unanticipated accidents. Moreover, mechanical equipment is often critical to the ability of a production process to perform as and when required [6]. Failure in such equipment can have serious safety, environment, and economic impact [7]. The aim of maintenance in mechanical equipment usually lies on preventing the equipment from failures and reduce maintenance costs by decreasing the number of unnecessary maintenance [8].

However, in most cases, the subsystems in mechanical machineries like bearings and gear transmission systems are not easily accessible, or hard to inspect visually the failures directly due to restrictions of time consuming disassembly, huge machine size, or environmental limitations [9]. Therefore, how to achieve early anomaly detection in mechanical equipment is always a hot issue in the field of mechanical maintenance [10]. The research target usually focuses on the identification of patterns in data that do not conform to expected behaviors. Many intelligent approaches for anomaly detection in mechanical equipment have been proposed and researched in the recent years. López-Pérez and Antonino-Daviu [11] applied infrared thermography to detect failure conditions in induction motors through comparing the temperature distribution between target and rather stable conditions. Lin et al. [12] proposed a crossover characteristics-based anomaly detection

approach by extracting failure features from nonlinear data in rotary machine. Griffin et al. [13] proposed a method based on neural networks and decision trees to detect anomalies for multiple machining processes, which demonstrates how intelligent control methods can detect different conditions in a robust and reliable manner. Lu et al. [14] introduced a stacked denoising autoencoder to distinguish anomaly and health condition of rotary machinery components. Aydin et al. [15] proposed a modified Kernel-based anomaly detection method to monitor the condition of catenary systems in a contactless manner. Li et al. [16] proposed a data-driven method based on deep belief networks to predict the time when anomalies may happen for maintenance scheduling in machining centers. In that research, deep learning algorithms has been be leveraged to predict potential failures. Peña et al. [17] used a rule-based system to detect anomalies in smart buildings from energy efficiency with a data mining approach. Zhou et al. [18] proposed a configurable method, which could support explicit knowledge representation with formal semantics and efficient knowledge utilization, for anomaly detection in machine tools. Diez-Olivan et al. [19] presented a method to detect anomaly conditions in monitoring sensor data using a kernel-based support vector machine. The proposed approach in that paper is interpretable and provides a tool for maintenance optimization based on real-time condition monitoring.

All these proposed methods have achieved certain targets in relevant domains and proved their feasibility in many cases. However, most of their methods inevitably require more or less information in stable condition or labels which could enable diagnostic models learn in a supervised manner. In many practical applications, a common dilemma is that the information about anomaly is not available or the history data is completely unlabelled [20]. To solve this dilemma, unsupervised learning methods have also been widely researched and successfully applied in many fields. Amruthnath and Gupta [21] applied several unsupervised learning methods including K-means, fuzzy C-means clustering etc. for early fault detection. In that paper [21], the authors leverage principal component analysis to simplify the environment and use elbow method and nbClust package to identify the number of clusters for unsupervised clustering in advance. Von Birgelen et al. [22] proposed a self-organizing maps inspired approach for unsupervised anomaly detection and location, in which the relationship between severity and degradation is provided by quantization error first, deviations from the diagnostic model could be then used to evaluate the severity degree. Costa et al. [23] proposed a self-developing fuzzy-rule-based classifier. In that paper, once the rules for classification were pre-set, the

classifier could automatically generate imaginary class labels according to the difference among input data. Serdio et al. [24] present a method for fault detection and identification in a power plant coal mills based on evolving fuzzy models and dynamic residual analysis, in which neither annotated samples nor fault patterns need to be priori available. In that study, the authors separated offline training stage from online detection phases to extract significant failure indicators in advance and detect anomaly in an unsupervised learning environment. However, it is still inevitable for most of those unsupervised approaches to require more or less certain empirical knowledge in corresponding domains to optimize number of clusters [21], demarcate degradation between normal condition and anomaly [22], pre-set initial rules for classification [23], or extract indicators from identified models [24]. In order to fill this gap, this paper proposes a novel deep learning-based method for anomaly detection in mechanical equipment through stacked autoencoders (SAE) and long short term memory (LSTM) neural networks. The proposed approach provides an alternative method to leverage and integrate features for anomaly detection instead of empirical knowledge such as number of clusters and calibrated degradation, especially when data are collected without labels. To validate our proposed method, we performed an experiment to detect anomaly condition for a rotating machine. According to the results, the proposed approach could detect anomaly condition with 99% accuracy under a completely unsupervised learning environment.

The remaining part of this paper is organized as follows. Section 2 introduces the set-up and data collection system of the experiment. Section 3 details SAE-based representation learning for multiple features sequence. Section 4 proposes a novel LSTM-based anomaly identification in time series along with the numerical results and discussion. Conclusions are summarized in the last section of this paper.

2. Set up and data collection

In the experiment, we used a Bently Nevada Rotor Kit RK3 to simulate the working condition of rotating equipment. Three acceleration meters of Kistler 8702B100 were mounted in three directions at the top of the bearing house to collect vibration signals in both normal and anomaly conditions. The test rig is shown in Figure 1. The collection frequency of the acceleration meters is 4096 Hz with the maximum revolving speed at 4000 rpm. Rub generator and mass adjustable load could be modulated to inject failures. The vibration

monitoring refers to a zero position of the test rig. Data recorded at the zero position would be recorded as the standard values in normal working condition.

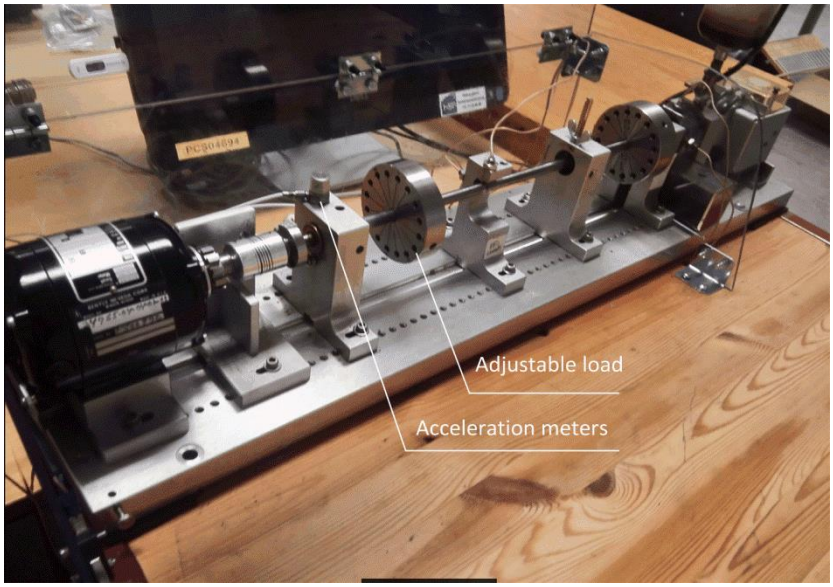

Fig. 1. Bently Nevada rotor kit

During the experiment, failures were injected through adding weights on the mass adjustable load to simulate load imbalance. The vibration signals will be measured through accelerometers at different rotating speed by means of proximity sensors and hand-held tachometer for control. The vibration signals in normal working condition will also be recorded and mixed together with unlabeled failure data. The label will only be used to validate and check the performance of proposed approach. In this research, we applied Wavelet Packet Decomposition (WPD) to extract wavelet coefficient-based and energy-based feature sequences from vibration signals to represent the working condition of target equipment in both time and frequency domains [25]. The essence of WPD is a wavelet transform where the discrete-time signal is parsed through more filters than the discrete wavelet transform, which can provide a multi-level time-frequency decomposition of signals [26]. Different types of wavelet functions may cause various time-frequency structures, in this paper, Daubechies 4 (DB4) wavelet function has been selected for the good performance in estimations of the local properties of signals like breakdown points [27], and the capacity to derive a set of conventional and energy-based features from signals [28]. After extracting features sequence from vibration signals, a common dilemma when analyzing vibration data from mechanical equipment is to determine the vibration level acceptance criteria through obtained multiple features sequence. It is also a challenge when using WPD to

extract failure information from multiple features sequence. In order to solve this challenge, SAE-based representation learning and LSTM-based anomaly identification will be introduced and applied to analyze vibration data in the following sections.

3. SAE-based representation learning for multiple features sequence

As mentioned above, in many practical applications, one may face the dilemma that the history data is collected and recoded unlabeled, let alone classified. To solve this challenge, we propose a SAE-LSTM anomaly detection method in this paper to identify the anomaly condition in an unsupervised learning environment.

When the history data is collected without labels (These labels usually can be used to represent the working condition in a supervised learning manner), an alternative method is to track the changes in multiple features sequence with time-series to identify the anomaly. As shown in Figure 2, features in time domain, frequency domain, or time-frequency domain were first extracted from vibration signals. The original data includes 33 features, which are extracted through WPD and Fourier Transform from vibration signals, in both time and frequency domains. Features 1-6 represent the peak and the second peak of vibration signals at X, Y, and Z directions in frequency domain after Fourier Transform. Features 7-18 denote the standard deviation noises of wavelet coefficients at level 1-4 in direction X, Y, and Z. Features 19-33 express wavelet packet energy features. To be more specific, Features 19-21 are the percentages of energy corresponding to the approximation in three directions and Features 22-33 denote the percentages of energy corresponding to the details at level 1-4 and three directions.

After unity-based normalization, all the extracted features would be transformed into the same scale. To prevent the inputs from explosion, SAE-based representation learning will further be leveraged to reduce the number of features extracted from raw data, and reconstruct the multiple features sequence.

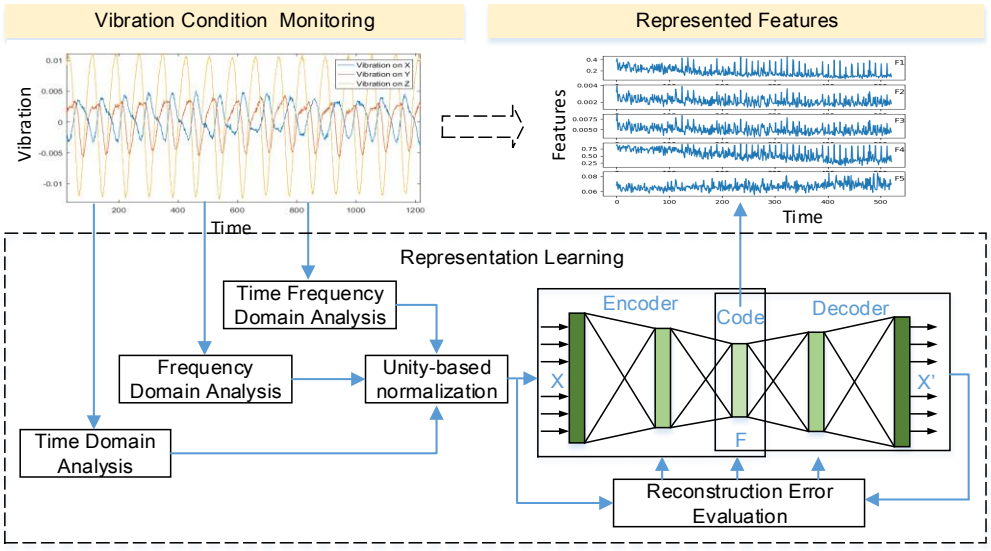

Fig. 2. Process of SAE-based representation learning for multiple features sequence

3.1 Feature normalization

During the experiment, vibration signals in both normal and anomalous conditions are collected through three accelerometers (Kistler 8702B100) at the sampling frequency of 4096 Hz. Each sampling unit has been divided into 10 parts with the same length for time series-based detection. Features of each part are first extracted through WPD and Fourier Transform, and will further be represented by autoencoders to reduce the dimension of features sequence. To adjust values measured on different scales to a notionally common scale, unity-based normalization [29] has been applied to normalize the inputs F'_{ij} for SAE-based representation learning, as shown in Equation 1.

$$F'_{ij} = \frac{F_{ij} - F_i^{min}}{F_i^{max} - F_i^{min}} \tag{1}$$

Where F_{ij} denotes the i^{th} feature in j^{th} samples, F_i^{min} and F_i^{max} represent the minimum and maximum values of the i^{th} feature in database, respectively. Figure 3 shows part of energy-based features after normalization, which will be used as inputs for SAE-based dimension reduction (Anomaly sampling units are collected when failures are injected, but the labels will be used in validation only). Visually, after normalization, data collected in anomaly still keep certain divergence from normal condition, though we cannot catch the rules directly in this step. It should be noticed that, during the training process, it is supposed

that we only have the data in normal condition. Data in anomaly is only collected to test and validate the proposed method.

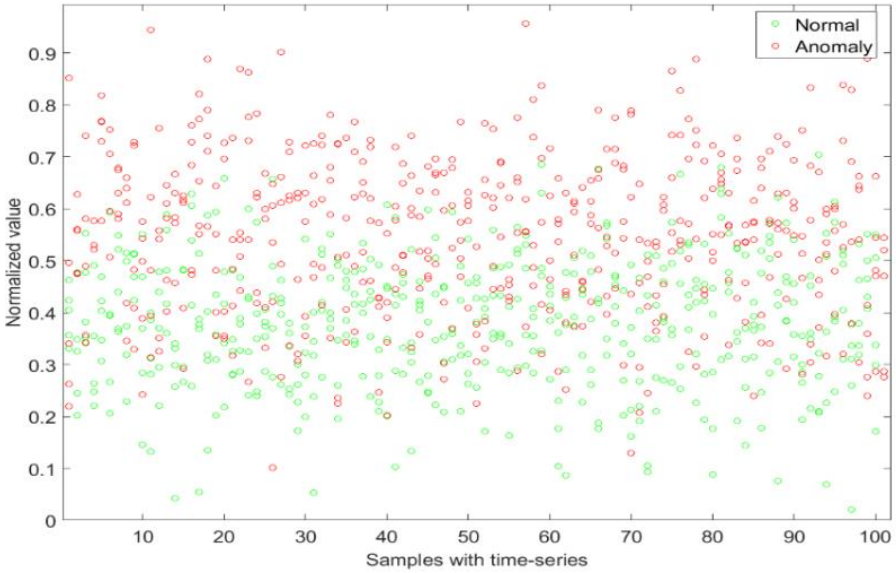

Fig. 3. Part of energy-based features after normalization

3.2 Feature representation with SAE

After feature normalization, sparse autoencoders have been leveraged to construct the deep neural network for representation learning. SAE is first proposed in 2007 [30,31]. It is a special type of deep neural networks created through stacking multiple autoencoder layers. The architecture of the deep neural network is pre-trained through single autoencoder layer by layer [32]. The output of SAE is the data input itself, which is leveraged for learning efficient encoding or dimensionality reduction for a set of data. More specifically, it is a nonlinear feature extraction method involving no class labels. Hence, it is generative. When an autoencoder uses three or more layers in the neural network, and the number of hidden layers is greater than one, the autoencoder is considered to be deep [33].

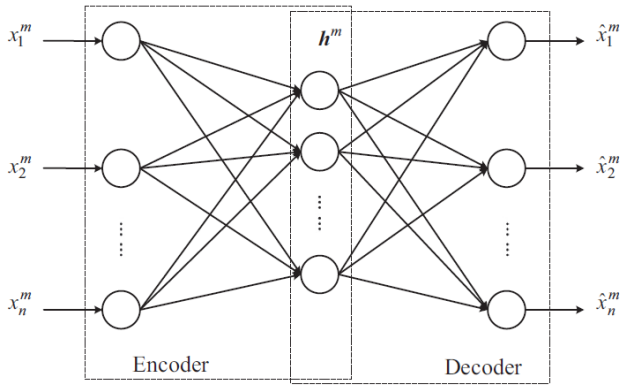

Fig. 4. Architecture of an autoencoder

As depicted in Figure 4, the input layer and hidden layer construct the encoder network, which transforms the input data from a high-dimensional space into codes as a low-dimensional space and the decoder network, which consists of the hidden layers and output layer, reconstructs the inputs from the corresponding codes. The encoder network is explicitly defined as an encoding function denoted by f_θ, which is also called as the encoder [32]. For each input signal $\boldsymbol{x}^m$ from a dataset $\{\boldsymbol{x}^m\}_{m=1}^M$ (M is the number of training samples), we label h^m as the obtained encode vector:

$$\boldsymbol{h}^m = f_{\boldsymbol{\theta}}(\boldsymbol{x}^m) \tag{2}$$

The decoder network is defined as a reconstruction function denoted by $g_{\theta'}$, namely the decoder. It maps $\boldsymbol{h}^m$ from the low-dimensional space back into the high-dimensional space, producing a reconstruction as Equation 3:

$$\hat{\boldsymbol{x}}^m = g_{\boldsymbol{\theta}'}(\boldsymbol{h}^m) \tag{3}$$

The parameter sets of the encoder and decoder are learned simultaneously on the task of reconstructing as well as possible the original input, attempting to incur the lowest possible reconstruction error $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$ over the M training samples. $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$is a loss function that measures the discrepancy between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ [32].

In summary, the autoencoder training aims to find the parameter sets $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ minimizing reconstruction error, which can be depicted as Equation (4):

$$\varphi_{AE}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{M}\sum_{m=1}^{M} L(\boldsymbol{x}^m, g_{\boldsymbol{\theta}'}(f_{\boldsymbol{\theta}}(\boldsymbol{x}^m))) \tag{4}$$

A deep neural network could be constructed by stacking multiple autoencoder layers with a final classification or regression layer on top. Stacking multiple autoencoder layers together allows the network to learn higher order features, where each successive layer represents additional complexity within the input

data [34]. Since each hidden layer in SAE is pre-trained through learning multiple nonlinear transformation of the inputs indecently, SAE has the ability to capture the main variations, discover the discriminative information, and represent the features from the raw data [35]. For predictive maintenance, representations of working condition with lower-dimension can improve performance in many situations such as fault classification and detection, especially when the input data is industrial big and row data. With the code vector of the previous trained autoencoder as input for training the next autoencoder, SAE could recognize the characteristics and effectively discover the discriminative information of these signals [36], and subsequently represent mechanical health conditions in a feasible and representational manner. Some practical applications have also shown that SAE has the ability to automatically mine the important information from the frequency spectra according to the diagnosis issues [14]. For this reason, we applied SAE as a representation learning model to reduce the dimension of original features sequence. The SAE constructed during the test has three hidden layers trained through L2 regularization in an unsupervised learning manner with the hidden layers of size 20, 10, and 5, respectively, and a linear transfer function for the decoder. The L2 weight regularization, sparsity regularization, and sparsity proportion have been set to 0.001, 4, and 0.05, respectively. After representation learning, the original 33 features are transformed into a multiple features sequence with time series, as shown in Figure 5.

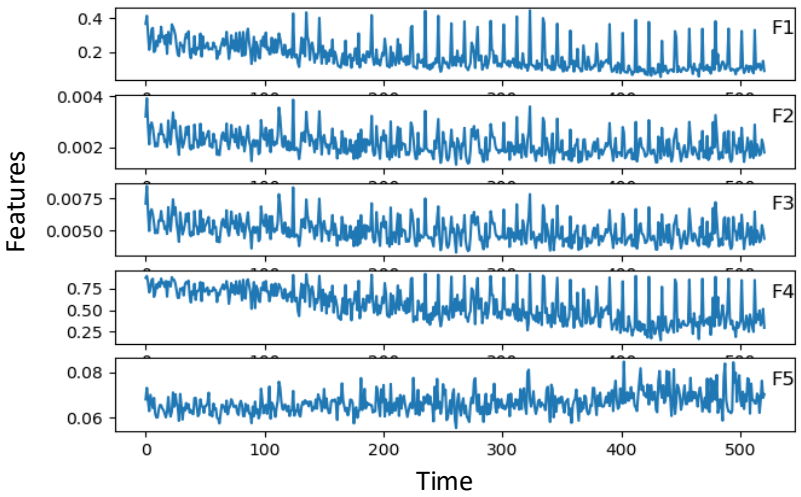

Fig. 5. Features after representation learning

4. LSTM-based anomaly identification with time series

4.1 Architecture of LSTM

To solve the problem of gradient vanishing in deep learning, a common method is to employ special architectures unaffected by gradient. LSTM is the most typical model in this type of deep learning architectures, which could avoid the fundamental problem of gradients vanishing through special architectures [37]. LSTM neural network is a type of recurrent neural network (RNN) proposed in 1997 to address the problem of insufficient, decaying error backflow in RNN training [38]. The basic idea of LSTM is simple: In a LSTM neural network, memory cells are employed as independent activation functions and identity functions with fixed weights, which are connected to themselves. Due to fixed weight, errors back-propagated through a memory cell cannot vanish or explode but stay as they are [39]. The weight matrixes in conventional RNNs are also trained via backpropagation through time series like the training process of normal neural network. Therefore, the gradients vanishing problem also happens in RNN while the complexity of the network increases, which means traditional RNN do not have the ability to discover information or capture dependencies hidden in long-term time series. In this background, LSTM was proposed to prevent back-propagated errors from gradients vanishing or exploding in RNN to deal with issues about long-term dependencies. The core idea behind the LSTM architecture is a memory cell, which can maintain its state over time, and non-linear gating units regulating the information flow into and out of the cell [40]. Compared with traditional RNN, LSTM neural network leverages memory cells with forget gates instead of traditional neurons to establish connections between inputs and outputs [41]. These adopted forget gates can effectively control the utilization of information in the cell states, and enable LSTM the capability to capture nonlinear dynamics in time series sensory data and learn effective representation of machine [42]. As shown in Figure 6, LSTM applies four special and interacting neural network layers, layer α, β, γ , o , instead of a single layer as in a standard RNN [43]. The first layer α is a sigmoid layer also called as forget gate layer, which returns a value between 0 and 1 in the previous cell state C_{t-1} , while 0 means no information pass and 1 means all information pass. The equation of the first layer can be denoted as Equation (5).

$$\alpha_t = \sigma(W_\alpha \cdot [h_{t-1}, x_t] + b_\alpha) \tag{5}$$

In Equation (5) – (10), σ is the sigmoid function, W_α is the weight of layer α, [ ] denotes the concatenate operation, x_t is the input x and time t, h_t is the output with respect to x_t, W_α , W_β , W_γ , W_o are the weights and b_α , b_β , b_γ, b_o are the biases of the layer α, β, γ, o, respectively. The second layer β is called as input gate layer, which is applied to decide which value shall be updated, denoted as Equation (6)

$$\beta_t = \sigma\big(W_\beta \cdot [h_{t-1}, x_t] + b_\beta\big) \tag{6}$$

Next, a tanh layer γ updates the values to be stored using:

$$\gamma_t = tanh\big(W_\gamma \cdot [h_{t-1}, x_t] + b_\gamma\big) \tag{7}$$

Where $tanh$ is the hyperbolic tangent function.

Then, we can update the previous state C_{t-1} to the current state C_t by Equation (8)

$$C_t = \alpha_t \cdot C_{t-1} + \beta_t \cdot \gamma_t \tag{8}$$

The final layer is also a sigmoid function layer, which determines what parts of the cell state will be the output, as denoted by Equation (9)

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

Then, the cell state go through tanh function and form the final output as Equation (10)

$$h_t = o_t tanh(C_t) \tag{10}$$

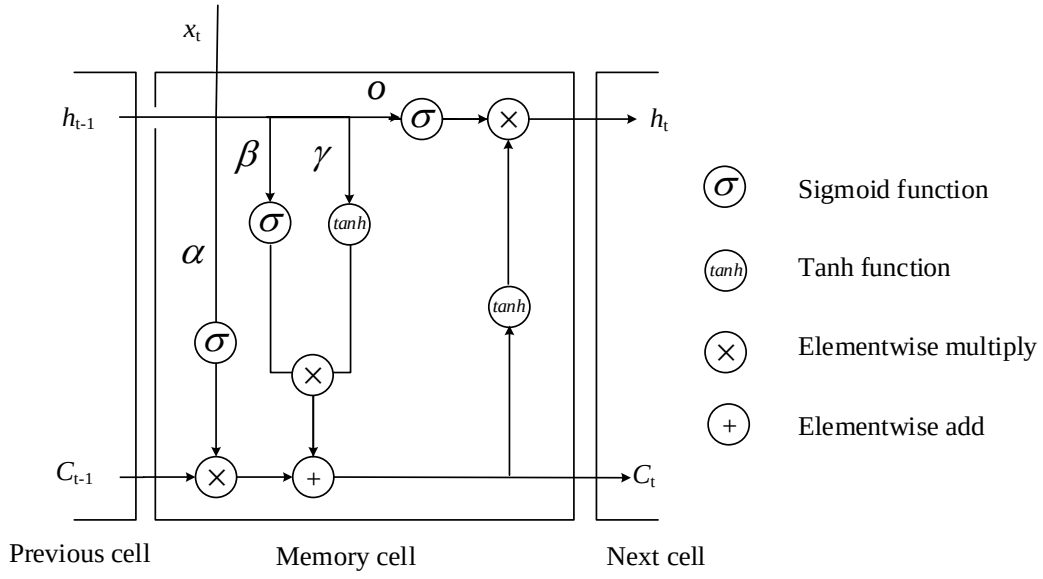

Fig. 6. Memory cell in LSTM

Through stacking multiple memory cells on top of each other, deep RNN can be created with the output sequence of one layer that forms the input sequence for the next, which then enables LSTM to discover

information from a dynamically changing contextual window over the input sequence history, rather than a static one as in the fixed-sized window applied in feed-forward neural networks [44].

4.2 Application of LSTM in anomaly identification

To deal with issues with high temporal dependency, a RNN is a natural choice due to the recurrent connections in the network, which allows the network to store memories of past information [45]. However, as discussed in the last section, standard RNN does not have the ability to learn long-term time dependencies because of the gradient vanishing problem. LSTM can solve this fundamental problem by applying the special memory cells in the architecture [46]. By stacking memory cells, information of previous inputs can be kept in the output to some degree, carried by cell state, which makes LSTM an outstanding tool to mimic time series [47]. This is the reason we would introduce LSTM as the prediction model with time series for the proposed anomaly detection approach. The LSTM network leveraged in our experiment is constructed in python environment with Keras deep learning library running on top of TensorFlow library developed by Google. During the experiment, each sampling unit of raw vibration signals is divided into 10 parts before feature extraction. Therefore, the LSTM model is constructed to predict the 10^{th} parts through the previous 9 parts. Each step includes five features in length. During the experiment, the multiple features sequence, which is obtained through SAE-based representation learning and unity-based normalization, is leveraged as the inputs of LSTM neural network. Figure 7 illustrates the process of our proposed SAE-LSTM approach for anomaly detection. In this case, the number of steps N is up to 10. The raw vibration signals were first divided into 10 parts. After SAE-based representation learning, the features at first 9 steps in each feature sequence will be used as inputs to map the features at 10^{th} step during the training process. Therefore, the applied LSTM neural network is constructed with 9 LSTM memory cells to represent the previous 9 steps in multiple features sequence and to predict the features at 10^{th} step. The error between predicted and actual values of the features at 10^{th} step will be leveraged to determine whether the equipment works in a normal condition.

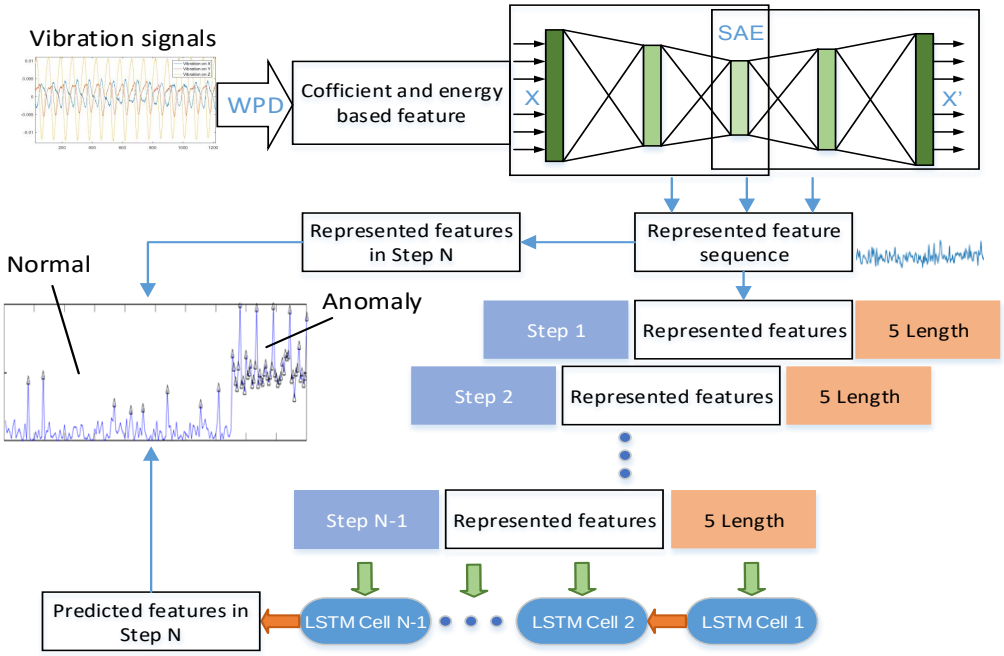

Fig. 7. Process of SAE-LSTM approach for anomaly detection

During the experiment, a selection of 500 samples is applied to train the LSTM neural network with fivefold cross-validation to validate the proposed approach. Figure 8 illustrates the numerical result of fivefold cross-validation, including the mean square errors of all the features and their average values.

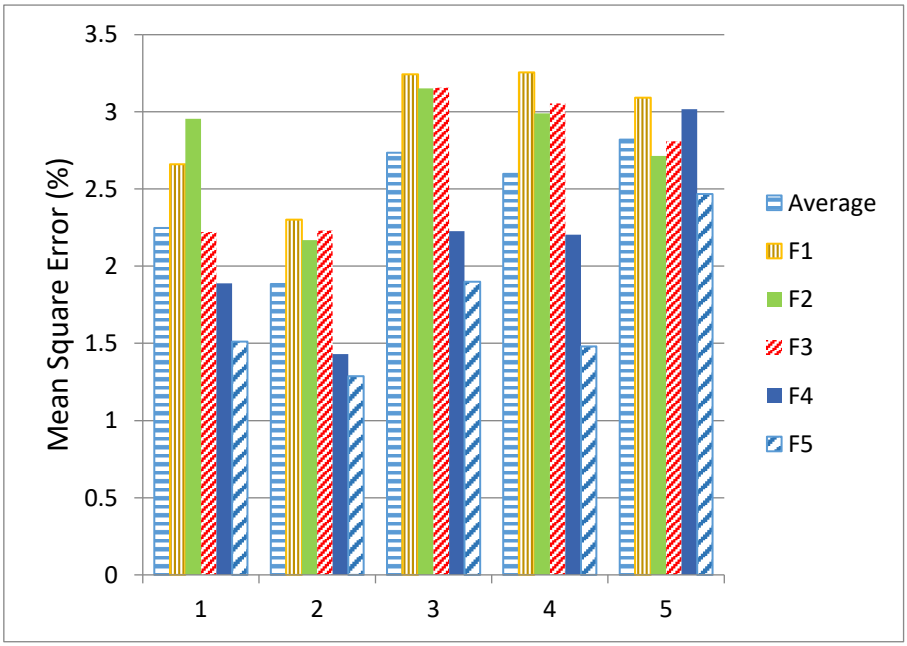

Fig. 8. Numerical result of 5-fold cross-validation

The numerical result of 5-fold cross validation shows that the proposed SAE-LSTM approach has the ability to predict multiple features sequence. The mean square errors of all the features applied during cross validation are below 3.5%. In addition, the diversities of the same feature are below 1%, which validates the stability of the proposed method. Since the second run during the process of cross validation shares the best performance with lowest training error, we will leverage it to verify the performance of anomaly detection. Figure 9 shows the construction error with training epochs during the training process, in which the mean construction error started to converge at about the 300th epochs with tiny fluctuation. The average training errors of the LSTM neural network at all of the five feature sequences fluctuate between -0.4 to 0.3. Since the target is to distinguish the anomaly and normal working condition instead of predicting the multiple features sequence directly, the performance of proposed method needs to be further validated.

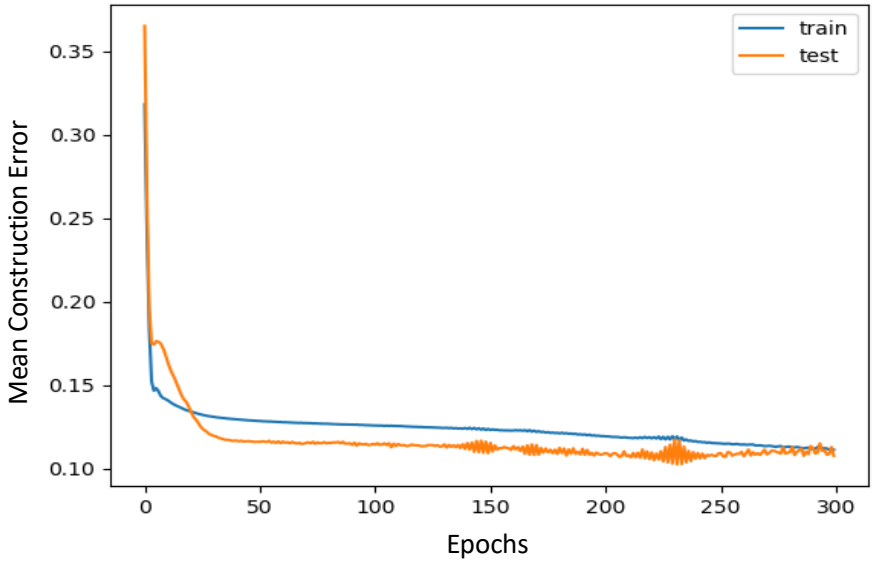

Fig. 9. Construction errors with training

4.3 Validation

To validate the performance of proposed SAE-LSTM approach for anomaly detection, a selection of 200 samples, constructed of 150 samples in normal condition and 50 samples in anomaly, is leveraged for testing. Figure 10 shows the testing result of anomaly detection through SAE-LSTM.

During the experiment, we applied the largest mean square error in each feature obtained through fivefold cross-validation [48] as the criterion to detect anomaly in the equipment. Since the sensitivity of each feature

to anomaly condition is highly subjective in nature, the criterion applied during the test is based on the overall performance in all features, which means only when all the prediction errors in five features are beyond the average values, the condition would be considered as anomaly. Table 1 lists the overall performance of proposed SAE-LSTM anomaly detection approach and the result of each single feature sequence, respectively.

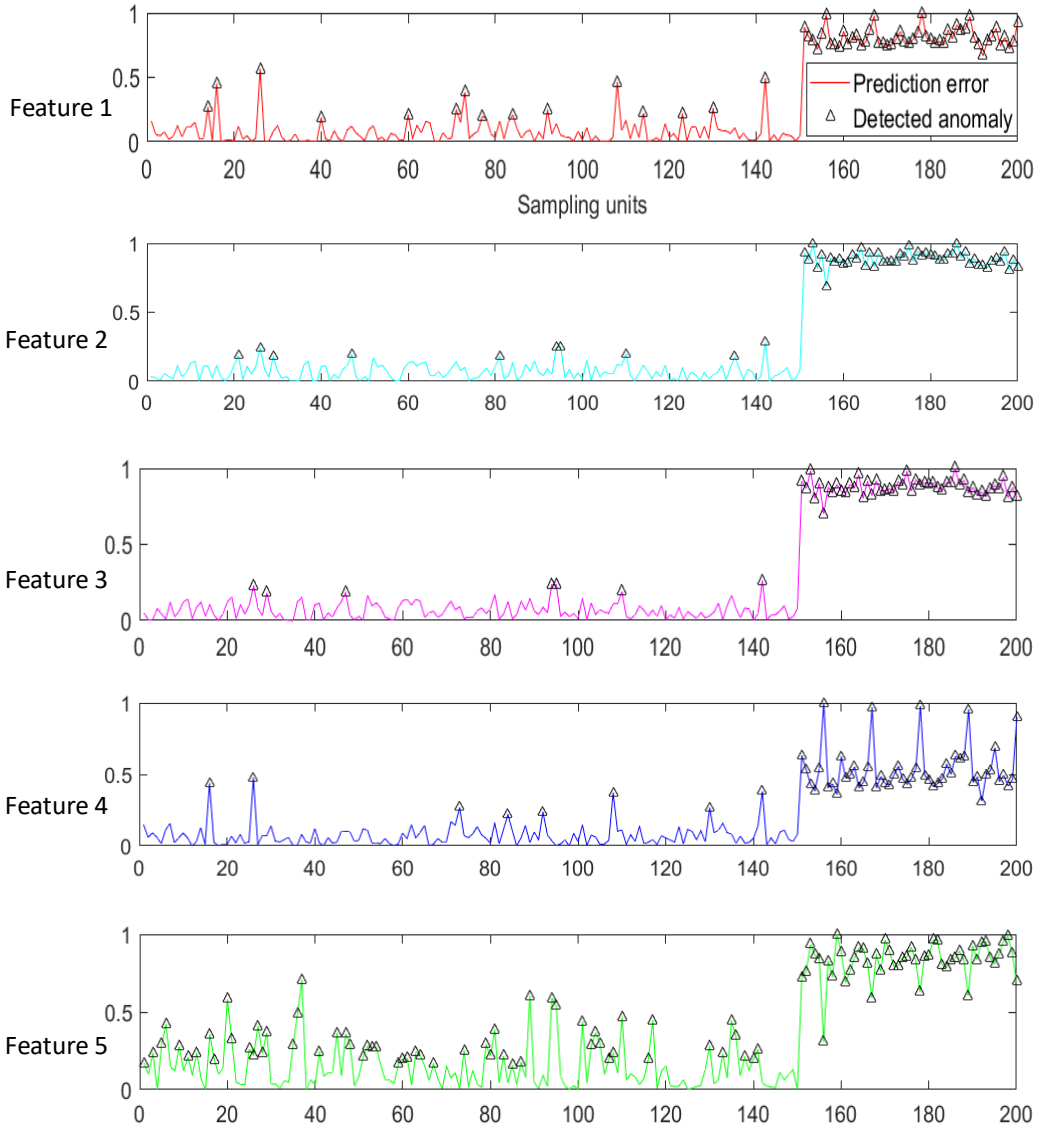

Fig. 10. Result of anomaly detection through SAE-LSTM

Table 1 Performance of SAE-LSTM for anomaly detection

Threshold (%)		Number of detected anomaly	Accuracy (%)
Feature 1	18.0418	65	92.5
Feature 2	17.7565	60	95

Feature 3	17.7676	57	96.5
Feature 4	17.3675	58	96
Feature 5	16.7939	109	71.5
F1&F2&F3&F4&F5		52	99

4.4 Discussion

The numerical results of each single feature sequence are shown in Table 1 together with the final result. As shown in Figure 10, there are some samples in normal condition misjudged by each single feature sequence individually. In addition, it is also obvious that not all the features trained through representation learning are suitable for anomaly detection (e.g., the accuracy of Feature 5 is only 71.5%). We consider the reason is that fault identification is a very subjective problem in nature, and the features after merging were extracted by SAE automatically with the distribution of original data, which makes the extracted features may be partly irrelevant or insensitive to the anomaly condition in this experiment. In this case, we assume that Feature 5 was trained to represent the original features, which are relatively irrelevant to the target, by SAE during the representation learning. However, we also want to highlight that it is also the superiority of deep learning methods since the features are extracted by machine automatically with certain rules hidden behind the distribution of original data instead of human experts with considerable empirical knowledge in the domain (the knowledge that may never be obtained in this case). It could also be a direction for future work to further figure out and testify the principles or rules about how to select the number of learning characteristics and parameters to improve the performance of representation learning without ground truth.

After all the 5 features are combined as the criteria to evaluate the working condition, which means anomaly samples shall be confirmed only after being identified by all the features, the overall detection accuracy could achieve 99%, which validate the performance of proposed SAE-LSTM method for anomaly detection. In this research, the data-driven model was trained and validated in a completely unsupervised learning environment, which means the proposed SAE-LSTM approach could ideally detect anomaly working condition when the data is collected without labels. In practical applications, if part of the data is collected with labels, it may help to optimize the detection criterion and further improve the detection accuracy, which would be a direction for future research.

In our previous work [49], we also did some research based on the same conditions with some supervised methods for fault classification and degradation assessment in a supervised learning environment, in which all the data-driven models are trained with labels or ground truth at the very beginning. The samples used in that research have been divided into 4 groups, normal condition and three types of injected failures, and the correct classification rate of applied back-propagation neural network, support vector machine, K-nearest neighbor classification, deep belief neural network, and fully connected deep neural network are 99.77%, 99.85%, 99.85%, 99.8%, and 99.87%, respectively. As the following research of that paper, samples collected in one type of the failures are merged with the normal condition without any labels to testify the feasibility of the proposed unsurprised anomaly detection approach in this paper. As mentioned above, after combining all the extracted features as criteria, the correct detection rate is 99%. From the perspective of Taskonomy [50], the target of machine learning in this research has been simplified into anomaly detection from classification. However, due to the missing of labels during training process, the difficulty of the task could also be considered as largely increased. Therefore, the reduction of detection accuracy in this paper is acceptable. Furthermore, with the development of automation and sensor technologies, the topic about how to analyze and leverage industry data without labels is increasingly significant in modern manufacturing industry. The proposed deep learning-based unsupervised method for anomaly detection could be a feasible solution to the issue. Future work could also focus on the recognition of different types of anomaly from unlabeled data.

5. Conclusion

This paper proposed a novel SAE-LSTM approach for anomaly detection in mechanical equipment. The proposed method could be divided into two stages: SAE-based multiple features sequence representation and LSTM-based anomaly identification. In order to validate and test the practical performance of proposed approach, an experiment for anomaly detection in rotary machinery through wavelet packet decomposition (WPD) and data-driven models is conducted. During the experiment, the results of fivefold cross-validation demonstrate the stability and performance of the proposed approach. The results also prove that the proposed SAE-LSTM approach could ideally detect anomaly working condition in a completely unsupervised learning environment through multiple features sequence when the history data is unlabeled and the empirical

knowledge about anomaly is absent. The proposed approach could provide alternative method to leverage and integrate features for fault diagnosis instead of empirical knowledge.

Funding information

The work described in this article has been conducted as part of the research project CIRCit (Circular Economy Integration in the Nordic Industry for Enhanced Sustainability and Competitiveness), which is part of the Nordic Green Growth Research and Innovation Programme (grant number: 83144), and funded by NordForsk, Nordic Energy Research, and Nordic Innovation.

References

1. Lei, Y., Lin, J., He, Z., & Zuo, M. J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 35(1), 108-126.
2. Hernandez-Vargas, M., Cabal-Yepez, E., & Garcia-Perez, A. (2014). Real-time SVD-based detection of multiple combined faults in induction motors. *Computers & Electrical Engineering*, 40(7), 2193-2203.
3. El Kadiri, S., Grabot, B., Thoben, K.-D., Hribernik, K., Emmanouilidis, C., von Cieminski, G., et al. (2016). Current trends on ICT technologies for enterprise information systems. *Computers in Industry*, 79(Supplement C), 14-33. doi:https://doi.org/10.1016/j.compind.2015.06.008.
4. Precup, R.-E., Angelov, P., Costa, B. S. J., & Sayed-Mouchaweh, M. (2015). An overview on fault diagnosis and nature-inspired optimal control of industrial process applications. *Computers in Industry*, 74(Supplement C), 75-94. doi:https://doi.org/10.1016/j.compind.2015.03.001.
5. Wang, P., & Guo, C. (2013). Based on the coal mine's essential safety management system of safety accident cause analysis. *American Journal of Environment, Energy and Power Research*, 1(3), 62-68.
6. Ayele, Y. Z., & Barabadi, A. (2016) 'Risk based inspection of offshore topsides static mechanical equipment in Arctic conditions' *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 4-7 Dec. 2016. pp. 501-506.
7. Gao, Z., Cecati, C., & Ding, S. X. (2015). A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. *IEEE Transactions on Industrial Electronics*, 62(6), 3757-3767. doi:10.1109/TIE.2015.2417501.
8. Klingert, F., Roeder, G., Schellenberger, M., Bauer, A., Frey, L., Brueggemann, M., et al. (2017) 'Condition-based maintenance of mechanical setup in aluminum wire bonding equipment by data mining' *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on*. IEEE, pp. 72-77.
9. Yang, Y., Dong, X., Peng, Z., Zhang, W., & Meng, G. (2015). Vibration signal analysis using parameterized time–frequency method for features extraction of varying-speed rotary machinery. *Journal of Sound and Vibration*, 335, 350-366.
10. Bangalore, P., & Tjernberg, L. B. (2015). An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings. *IEEE Transactions on Smart Grid*, 6(2), 980-987. doi:10.1109/TSG.2014.2386305.
11. López-Pérez, D., & Antonino-Daviu, J. (2017). Application of Infrared Thermography to Failure Detection in Industrial Induction Motors: Case Stories. *IEEE Transactions on Industry Applications*, 53(3), 1901-1908. doi:10.1109/TIA.2017.2655008.
12. Lin, J., & Chen, Q. (2014). A novel method for feature extraction using crossover characteristics of nonlinear data and its application to fault diagnosis of rotary machinery. *Mechanical Systems and Signal Processing*, 48(1), 174-187.
13. Griffin, J. M., Doberti, A. J., Hernández, V., Miranda, N. A., & Vélez, M. A. (2017). Multiple classification of the force and acceleration signals extracted during multiple machine processes: part 1 intelligent classification from an anomaly perspective. *The International Journal of*

Advanced Manufacturing Technology, 93(1), 811-823. doi:10.1007/s00170-017-0320-3.

14. Lu, C., Wang, Z.-Y., Qin, W.-L., & Ma, J. (2017). Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Processing*, 130, 377-388.
15. Aydin, I., Karakose, M., & Akin, E. (2015). Anomaly detection using a modified kernel-based tracking in the pantograph–catenary system. *Expert Systems with Applications*, 42(2), 938-948. doi:https://doi.org/10.1016/j.eswa.2014.08.026.
16. Li, Z., Wang, Y., & Wang, K. (2017). A data-driven method based on deep belief networks for backlash error prediction in machining centers. *Journal of Intelligent Manufacturing*. doi:10.1007/s10845-017-1380-9.
17. Peña, M., Biscarri, F., Guerrero, J. I., Monedero, I., & León, C. (2016). Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. *Expert Systems with Applications*, 56, 242-255. doi:https://doi.org/10.1016/j.eswa.2016.03.002.
18. Zhou, Q., Yan, P., Liu, H., Xin, Y., & Chen, Y. (2018). Research on a configurable method for fault diagnosis knowledge of machine tools and its application. *The International Journal of Advanced Manufacturing Technology*, 95(1), 937-960. doi:10.1007/s00170-017-1268-z.
19. Diez-Olivan, A., Pagan, J. A., Khoa, N. L. D., Sanz, R., & Sierra, B. (2018). Kernel-based support vector machines for automated health status assessment in monitoring sensor data. *The International Journal of Advanced Manufacturing Technology*, 95(1), 327-340. doi:10.1007/s00170-017-1204-2.
20. Landry, M., Leonard, F., Landry, C., Beauchemin, R., Turcotte, O., & Brikci, F. (2008). An Improved Vibration Analysis Algorithm as a Diagnostic Tool for Detecting Mechanical Anomalies on Power Circuit Breakers. *IEEE Transactions on Power Delivery*, 23(4), 1986-1994. doi:10.1109/TPWRD.2008.2002846.
21. Amruthnath, N., & Gupta, T. (2018) 'A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance' *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*. IEEE, pp. 355-361.
22. von Birgelen, A., Buratti, D., Mager, J., & Niggemann, O. (2018). Self-Organizing Maps for Anomaly Localization and Predictive Maintenance in Cyber-Physical Production Systems. *Procedia CIRP*, 72, 480-485. doi:https://doi.org/10.1016/j.procir.2018.03.150.
23. Costa, B. S. J., Angelov, P. P., & Guedes, L. A. (2015). Fully unsupervised fault detection and identification based on recursive density estimation and self-evolving cloud-based classifier. *Neurocomputing*, 150, 289-303. doi:https://doi.org/10.1016/j.neucom.2014.05.086.
24. Serdio, F., Lughofer, E., Pichler, K., Buchegger, T., & Efendic, H. (2014). Residual-based fault detection using soft computing techniques for condition monitoring at rolling mills. *Information Sciences*, 259, 304-320. doi:https://doi.org/10.1016/j.ins.2013.06.045.
25. Hu, Q., He, Z., Zhang, Z., & Zi, Y. (2007). Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble. *Mechanical Systems and Signal Processing*, 21(2), 688-705. doi:https://doi.org/10.1016/j.ymssp.2006.01.007.
26. Zhang, Y., Liu, B., Ji, X., & Huang, D. J. N. P. L. (2017). Classification of EEG Signals Based on Autoregressive Model and Wavelet Packet Decomposition (journal article). 45(2), 365-378. doi:10.1007/s11063-016-9530-1.
27. Ferreira, C. B. R., & Borges, D. b. L. (2003). Analysis of mammogram classification using a wavelet transform decomposition. *Pattern Recognition Letters*, 24(7), 973-982.
28. Murugappan, M., Ramachandran, N., & Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 3(04), 390.
29. Rastbood, A., Majdi, A., & Gholipour, Y. (2017). Prediction of structural forces of segmental tunnel lining using FEM based artificial neural network. *Int. Journal of Mining & Geo-Engineering*, 51(1), 71-78.
30. Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007) 'Greedy layer-wise training of deep networks' *Advances in neural information processing systems*. pp. 153-160.
31. Poultney, C., Chopra, S., & Cun, Y. L. (2007) 'Efficient learning of sparse representations with an energy-based model' *Advances in neural information processing systems*. pp. 1137-1144.
32. 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.
33. Deng, L. (2012). Three classes of deep learning architectures and their applications: a tutorial survey. *APSIPA transactions on signal and*

information processing.

34. Galloway, G. S., Catterson, V. M., Fay, T., Robb, A., & Love, C. (2016). Diagnosis of tidal turbine vibration data through deep neural networks. *In: Proceedings of the Third European Conference of the Prognostics and Health Management Society 2016.* , (PHM Society), 172-180.
35. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb), 625-660.
36. Jiang, S., Chin, K.-S., Wang, L., Qu, G., & Tsui, K. L. (2017). Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. *Expert Systems with Applications*, 82, 216-230. doi:https://doi.org/10.1016/j.eswa.2017.04.017.
37. Cortez, B., Carrera, B., Kim, Y.-J., & Jung, J.-Y. (2018). An architecture for emergency event prediction using LSTM recurrent neural networks. *Expert Systems with Applications*, 97, 315-324. doi:https://doi.org/10.1016/j.eswa.2017.12.037.
38. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735.
39. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
40. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*.
41. Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54, 187-197.
42. Zhao, R., Yan, R., Wang, J., & Mao, K. (2017). Learning to monitor machine health with convolutional bi-directional lstm networks. *Sensors*, 17(2), 273.
43. Liao, L., & Ahn, H.-i. (2016). Combining Deep Learning and Survival Analysis for Asset Health Management. *International journal of prognostics and health management*.
44. Sak, H., Senior, A., & Beaufays, F. (2014) 'Long short-term memory recurrent neural network architectures for large scale acoustic modeling' *Fifteenth Annual Conference of the International Speech Communication Association*.
45. Sutskever, I., Vinyals, O., & Le, Q. V. (2014) 'Sequence to sequence learning with neural networks' *Advances in neural information processing systems*. pp. 3104-3112.
46. de Bruin, T., Verbert, K., & Babuška, R. (2017). Railway track circuit fault diagnosis using recurrent neural networks. *IEEE transactions on neural networks and learning systems*, 28(3), 523-533.
47. Zhuge, Q., Xu, L., & Zhang, G. (2017). LSTM Neural Network with Emotional Analysis for Prediction of Stock Price. *Engineering Letters*, 25(2), 167-175.
48. Cheng, S., & Pecht, M. (2012). Using cross-validation for model parameter selection of sequential probability ratio test. *Expert Systems with Applications*, 39(9), 8467-8473. doi:https://doi.org/10.1016/j.eswa.2012.01.172.
49. Li, Z., Wang, Y., & Wang, K. (2019). A deep learning driven method for fault classification and degradation assessment in mechanical equipment. *Computers in Industry*, 104, 1-10. doi:https://doi.org/10.1016/j.compind.2018.07.002.
50. Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018) 'Taskonomy: Disentangling task transfer learning' *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3712-3722.