Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

Plug-and-Play Unsupervised Fault Detection and Diagnosis for Complex Industrial Monitoring

Maksim Golyadkin[1,4] , **Maria Shtark**[2] , **Petr Ivanov**[2] , **Alexandr Kozhevnikov**[2] , **Leonid Zhukov**[3] , **Ilya Makarov**[4,2,5]
[1]HSE University
[2]AI Talent Hub, ITMO University
[3]BCG X
[4]AIRI
[5]ISP RAS Research Center for Trusted Artificial Intelligence
golyadkin@airi.net

Abstract

Today industrial facilities are equipped with lots of sensors throughout all the production line for monitoring means. Gathered data can be used to detect and predict failures; however, manual labeling of large amounts of data for supervised learning is complicated. This paper introduces an innovative approach to unsupervised fault detection and diagnosis tailored for monitoring industrial chemical processes. We showcase the efficacy of our model using two publicly accessible datasets from the Tennessee Eastman Process, each containing various faults. Furthermore, we illustrate that by fine-tuning the model on a limited amount of labeled data, it achieves performance close to that of a state-of-the-art model trained on the entire dataset.

1 Introduction

Specialized equipment and technology are utilized in chemical processing plants to facilitate the manufacturing process. A closed-loop control system is commonly employed to ensure the stability of production by automatically making small adjustments. Process lines are extensively instrumented with sensors that provide real-time data to monitoring and control systems. Despite these measures, unexpected process deviations can sometimes occur, resulting in reduced process yield, interruptions, and potential equipment damage or failure.

Early fault detection and diagnostics are crucial components of a process monitoring system. A fault is commonly defined as a deviation of a process variable outside the acceptable production range [Venkatasubramanian *et al.*, 2003]. Detecting faults early enables the selection of appropriate recovery procedures to restore the process to its normal operating state.

Fault detection and diagnostics (FDD) techniques typically use a supervised approach, where you have to label all sensor data for each time interval with the corresponding process state. However, manual labeling of extensive data sets can be costly and challenging in industrial environments. In response to this issue, unsupervised FDD methods have been put forward as an alternative approach. These unsupervised methods reduce the dimensionality of sensor data and utilize clustering techniques to group samples based on their process states.

Recently, various deep clustering techniques have been proposed for unsupervised image classification [Ji *et al.*, 2019; Hu *et al.*, 2017; Huang *et al.*, 2020; Li *et al.*, 2021; Grechikhin and Savchenko, 2019; Savchenko *et al.*, 2022; Niu *et al.*, 2020]. Nevertheless, these approaches utilize feature extractors that primarily focus on low-level features, overlooking the potential high-level characteristics hidden within the input data.

We introduce a new unsupervised FDD method that utilizes deep learning techniques to achieve high accuracy with chemical sensor data. Our method combines SSL pretraining and Deep Clustering to enhance performance in this domain.

Our approach outperforms contemporary unsupervised FDD methods when evaluated using multiple clustering metrics on the Tennessee Eastman Process (TEP) benchmarks.

2 Related Work

In recent times, there has been a surge in data-driven techniques aimed at detecting and diagnosing faults in industrial processes [Lei *et al.*, 2020; Taqvi *et al.*, 2021; Bi *et al.*, 2022; Lomov *et al.*, 2021a; Pozdnyakov *et al.*, 2024]. The majority of these methods rely on extracting features from raw historical data. Nevertheless, data captured by industrial sensors often exhibit high redundancy and intercorrelation, posing challenges for many data-driven approaches.

Fault detection and diagnosis typically follow two broad approaches: supervised and unsupervised. In the supervised approach, a labeled dataset is necessary, enabling the trained model to differentiate between observed abnormal states in the process history. However, this dataset often lacks representation of all potential faults, prompting the exploration of unsupervised methods. Unlike supervised methods, unsupervised techniques are not limited to a predefined set of observed and labeled faults.

In the past few years there have been a lot of studies using modern techniques for fault detection and diagnosis problems: mostly deep neural networks of different types [Zhang and Qiu, 2022; Bi and Zhao, 2021; Golyadkin *et al.*, 2023; Lomov *et al.*, 2021a; Kovalenko *et al.*, 2022] and other.

2.1 Self-Supervised Learning for Time Series

Self-supervised learning techniques for time series processing can be divided into two main categories. The first group utilizes pretext tasks tailored to leverage the structural characteristics inherent in time series data. Conversely, methods within the second group adopt general representation learning strategies applicable to diverse data types, exhibiting superior performance across a spectrum of machine learning tasks in recent studies [Brown *et al.*, 2020; Dai *et al.*, 2021].

2.2 Deep Clustering

Basic deep clustering techniques rely on applying traditional clustering algorithms to features extracted using a pretrained feature extractor [Huang *et al.*, 2014; Xie *et al.*, 2016; Yang *et al.*, 2017; Li *et al.*, 2018]. These models can be trained using unlabeled data, operating by extracting semantically significant features from raw data.

Alternatively, there are models trained using end-to-end indirect loss functions to map inputs to cluster indices [Ji *et al.*, 2019; Hu *et al.*, 2017; Huang *et al.*, 2020; Li *et al.*, 2021; Niu *et al.*, 2020]. However, a notable drawback of this approach is the utilization of the same loss function for both feature extraction and class assignment training phases. Van Gansbeke *et al.* introduced the SCAN algorithm [Van Gansbeke *et al.*, 2020], which conducts clustering by ensuring the similarity of cluster predictions for neighboring objects, utilizing proximity calculated within the embedding space of the pretrained feature extractor.

3 Model Description

Our method aims to generate clustering results for a given set of unlabeled multivariate time series samples $\mathcal{X} = X_1, \ldots, X_N$, where each X_i is a matrix in $\mathbb{R}^{L \times D}$ denoting sample length L and sensor count D. Corresponding to $\mathcal{X}$, we have a set of labels $Y = y_1, \ldots, y_N$, where $y_i \in 1, \ldots, Q$, but these labels are unavailable during training. Our model comprises two components: a large feature extractor $\mathcal{F}$ and a small clustering network $\mathcal{C}$. Initially, the feature extractor undergoes pretraining using self-supervised learning methods, and the cluster count is determined through visual analysis of the latent space distribution. Subsequently, the clustering network and feature extractor are jointly trained using the SCAN loss function [Van Gansbeke *et al.*, 2020], using information from the embedding space regarding neighboring data points.

Self-supervised learning techniques are employed to empower our model in discerning distinct samples without relying on ground truth labels. Through these methodologies, the neural network delves into the intrinsic structure of the data. We utilize a feature extractor built upon the Transformer architecture [Vaswani *et al.*, 2017], which comprises four components: an encoder $\mathcal{T}$, a sequential pooling layer $\mathcal{P}$, a projection head $\mathcal{H}$, and a reconstruction head $\mathcal{R}$.

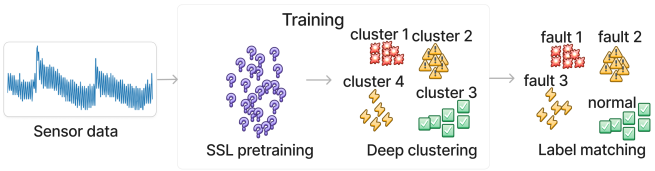

Figure 1: Overview of our method.

To create masked samples from a given sample $X \in \mathbb{R}^{L \times D}$, we generate a binary mask $M \in 0, 1^{L \times D}$ and perform element-wise multiplication, resulting in $\hat{X} = X \odot M$. The loss function employed for this task is MSE, calculated exclusively for the masked values:

$$\mathcal{L}_{rec}(\hat{X}, X) = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{|M^i|} \sum_{l,d:\ M^i_{ld}=0} (\mathcal{R}(\mathcal{T}(\hat{X}))_{ld} - X_{ld})^2, \tag{1}$$

where B is the batch size, and l and d correspond to the number of timestamps and the number of sensors, respectively.

Training proceeds by minimizing the NT-Xent loss [Chen *et al.*, 2020] (normalized temperature-scaled cross entropy). Initially, a minibatch of B samples is randomly sampled and augmented to yield a minibatch of $2B$ samples. Consequently, for each sample, there exists one positive pair and $(2B - 2)$ negative pairs. The NT-Xent loss encourages the distance between positive pairs to be smaller than that between negative pairs. Specifically, for the positive pair (i, j), the loss is defined as follows:

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z_i}, \boldsymbol{z_j})/\tau)}{\sum k = 1^{2B} [k \neq i] \exp(\text{sim}(\boldsymbol{z_i}, \boldsymbol{z_k})/\tau)}, \tag{2}$$

where $\boldsymbol{z_i}, \boldsymbol{z_j}, \boldsymbol{z_k} \in \mathbb{R}^F$ represent the outputs of the projection head, $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / (||\mathbf{u}|| : ||\mathbf{v}||)$ denotes the cosine similarity, and τ denotes the temperature parameter. The final loss $\mathcal{L}_{cont}$ is computed across all $2B$ positive pairs.

The self-supervised learning (SSL) pretraining spans over E epochs, following this iterative process:

1. Randomly sample a minibatch of size B.
2. Apply weak augmentation α and strong augmentation β.
3. Independently generate a binary mask M_i for each sample.
4. Reconstruct the masked values and compute reconstruction loss.
5. Generate embedding for the masked samples and calculate the NT-Xent loss.
6. Compute total loss as a weighted sum of the two above.

Clustering organizes samples based on their similarity, effectively distinguishing various process states even without the presence of ground truth labels. Prior to training, a preprocessing step known as nearest neighbors mining is necessary. For each sample X_i, we identify its K nearest neighbors $\mathcal{N}_{X_i}$ within the embedding space of the feature extractor.

The nearest neighbors mining algorithm is as follows:

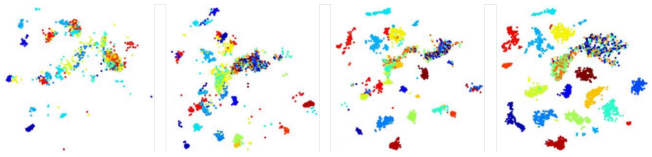

Figure 2: Evolution of clusters during our model training.

1. Randomly shuffle the training dataset.
2. Split the dataset into T chunks of equal size.
3. For each X_i, the nearest neighbors are found within the chunk $\boldsymbol{X}_k$ it belongs to: $\mathcal{N}_{X_i} = \text{NearestNeighbors}(X_i, \boldsymbol{X}_k)$.

The clustering network $\mathcal{C}$ is structured as a 2-layer Multi-Layer Perceptron (MLP), with intermediate BatchNormalization and ReLU activation function, and final softmax activation, $\mathcal{C}\colon \mathbb{R}^F \to \mathbb{R}^{\tilde{M}}$. The learning mechanism is engineered to ensure consistent label prediction distributions for both the sample and its neighboring instances. In sum, the loss computation proceeds as follows:

$$\mathcal{L}_{SCAN} = -\frac{1}{B}\sum_{i=1}^{B} \log \langle \mathcal{C}(\mathcal{F}(X_i)), \mathcal{C}(\mathcal{F}(X_i^{NN}))\rangle + \quad (3)$$

$$+\lambda_{ent}\, H(\mathcal{C}'), \quad \text{where } \mathcal{C}' = \frac{1}{B}\sum_{i=1}^{B} \mathcal{C}(\mathcal{F}(X_i)), \quad (4)$$

and $\langle \cdot, \cdot \rangle$ denotes the dot product, λ_{ent} is entropy loss weight, X_i^{NN} is the neighbour randomly sampled from $\mathcal{N}_{X_i}$, and $H(\mathcal{C}')$ denotes the entropy over discrete distribution $\mathcal{C}'$.

4 Tennessee Eastman Process

The Tennessee Eastman Process (TEP) is a widely recognized benchmark used to evaluate process control and FDD methods, originally developed by the Eastman Chemical Company. [Downs and Vogel, 1993]. There are two separate numerical simulators for the TEP process. An extended version of this dataset (hereafter $\text{TEP}_{\text{Rieth}}$) was first introduced and employed in [Rieth *et al.*, 2018]. The second simulator uses distinct control schemes.[Lawrence Ricker, 1996; Ricker and Lee, 1995; Larsson *et al.*, 2001] and is available at the Tennessee Eastman Challenge Archive. The extended dataset (hereafter $\text{TEP}_{\text{Ricker}}$) generated with this model is presented in [Reinartz *et al.*, 2021].

	PCA	ST-CatGAN	ConvAE	Ours
Detection TPR	0.36	0.30	0.48	**0.84**
Detection FPR	0.00	0.00	0.00	0.00
CDR	0.79	0.32	**0.93**	0.92
ADD	113.95	102.63	49.95	**5.21**

Table 1: Aggregated detection and diagnosis metrics evaluated on $\text{TEP}_{\text{Rieth}}$ in the unsupervised setting.

	PCA	ST-CatGAN	ConvAE	Ours
Detection TPR	0.36	0.36	0.64	**0.87**
Detection FPR	0.00	0.00	0.00	0.00
CDR	0.95	0.89	0.89	**0.96**
ADD	111.49	135.04	52.28	**28.47**

Table 2: Aggregated detection and diagnosis metrics evaluated on $\text{TEP}_{\text{Ricker}}$ in the unsupervised setting.

5 Evaluation Metrics

We conduct FDD on multivariate time series data $\mathcal{X} = X_1, \ldots, X_N$, where each X_i is a matrix in $\mathbb{R}^{L\times D}$ representing sample length L and sensor count D. Samples are generated using sliding windows of size L. In our experiments, window size is set to 100 and step size is set to 1. For datasets $\text{TEP}_{\text{Rieth}}$ and $\text{TEP}_{\text{Ricker}}$, this corresponds to 300 minutes and 3 minutes, respectively.

For FDD evaluation, we compare ground truth labels with predicted ones using label matching. Metrics include True Positive Rate (TPR), False Positive Rate (FPR), and Correct Diagnosis Rate (CDR). Detection TPR and FPR are calculated for each fault separately, while ADD measures the average number of samples between the first ground-truth faulty sample and the first detected faulty sample.

6 Experiments

We conducted a comparative analysis between our method and various unsupervised models, encompassing both traditional and state-of-the-art data-driven methods. We hypothesize that an effective unsupervised model should be capable of detecting all faults by partitioning samples into distinct groups, hence we set the number of clusters equal to the number of classes. Figure 1 shows the evolution of clusters during training. As baseline models, we utilized PCA [He *et al.*, 2005], ST-CatGAN [Tao *et al.*, 2020], ConvAE [Zheng and Zhao, 2020], and GRU [Lomov *et al.*, 2021b].

All four models on both $\text{TEP}_{\text{Rieth}}$ and $\text{TEP}_{\text{Ricker}}$ datasets have shown Detection FPR less than 0.01 which is below practical threshold.

7 Conclusion

The study presents an unsupervised FDD model utilizing self-supervised learning and deep clustering techniques, evaluated on $\text{TEP}_{\text{Rieth}}$ and $\text{TEP}_{\text{Ricker}}$ datasets. Our approach outperforms other unsupervised methods, attributed to the feature extractor's proficiency in encoding sensor data and deep clustering's ability to form dense clusters in the latent space.

We demonstrate the model's effectiveness with unlabeled data, common in industrial processes, and propose semi-supervised fine-tuning to detect challenging faults with minimal labeled data. Our method combines Self-Supervised Learning (SSL) with deep clustering, facilitating training on unlabeled data while considering the anticipated number of faults. Fine-tuning enables leveraging labeled examples within the dataset. Finally, adjusting the expected number of faults or incorporating new fault types allows fine-tuning of the feature extractor without retraining from scratch.

Ethical Statement

Though AI could be a great piece of automation, it shouldn't be blindly trusted to monitor and/or manage complex industrial systems. AI algorithms, even the most intelligent ones, may make critical mistakes leading to severe consequences: money and time loss, environmental or human harm, etc. This makes a huge field of study itself: intelligent algorithm monitoring, safe integration, and application.

No one should use these algorithms without a failsafe. In some cases, it's even better not to use AI-powered systems at all. We should always consider the whole picture: automation gain and possible losses due to its failure.

Acknowledgements

The work of I. Makarov on fault detection and giagnosis related work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of dated November 2, 2021 No. 70-2021-00142.

References

[Bi and Zhao, 2021] Xiaotian Bi and Jinsong Zhao. A novel orthogonal self-attentive variational autoencoder method for interpretable chemical process fault detection and identification. *Process Safety and Environmental Protection*, 156:581–597, 2021.

[Bi *et al.*, 2022] Xiaotian Bi, Ruoshi Qin, Deyang Wu, Shaodong Zheng, and Jinsong Zhao. One step forward for smart chemical process fault detection and diagnosis. *Computers I& Chemical Engineering*, 164:107884, 2022.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[Dai *et al.*, 2021] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[Downs and Vogel, 1993] James J Downs and Ernest F Vogel. A plant-wide industrial process control problem. *Computers & chemical engineering*, 17(3):245–255, 1993.

[Golyadkin *et al.*, 2023] Maksim Golyadkin, Vitaliy Pozdnyakov, Leonid Zhukov, and Ilya Makarov. Sensorscan: Self-supervised learning and deep clustering for fault diagnosis in chemical processes. *Artificial Intelligence*, 324:104012, 2023.

[Grechikhin and Savchenko, 2019] Ivan Grechikhin and Andrey V Savchenko. User modeling on mobile device based on facial clustering and object detection in photos and videos. In *Pattern Recognition and Image Analysis: 9th Iberian Conference, IbPRIA 2019, Madrid, Spain, July 1–4, 2019, Proceedings, Part II 9*, pages 429–440. Springer, 2019.

[He *et al.*, 2005] Q. Peter He, S. Joe Qin, and Jin Wang. A new fault diagnosis method using fault directions in fisher discriminant analysis. *AIChE Journal*, 51(2):555–571, 2005.

[Hu *et al.*, 2017] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1558–1567. JMLR.org, 2017.

[Huang *et al.*, 2014] Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. Deep embedding network for clustering. In *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537, 2014.

[Huang *et al.*, 2020] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8846–8855, 2020.

[Ji *et al.*, 2019] Xu Ji, Andrea Vedaldi, and Joao Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9864–9873, 2019.

[Kovalenko *et al.*, 2022] Alexander Kovalenko, Vitaliy Pozdnyakov, and Ilya Makarov. Graph neural networks with trainable adjacency matrices for fault diagnosis on multivariate sensor data, 2022.

[Larsson *et al.*, 2001] Truls Larsson, Kristin Hestetun, Espen Hovland, and Sigurd Skogestad. Self-optimizing control of a large-scale plant: The tennessee eastman process. *Industrial & Engineering Chemistry Research*, 40(22):4889–4901, 2001.

[Lawrence Ricker, 1996] N. Lawrence Ricker. Decentralized control of the tennessee eastman challenge process. *Journal of Process Control*, 6(4):205–221, 1996.

[Lei *et al.*, 2020] Yaguo Lei, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, and Asoke K. Nandi. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 138:106587, 2020.

[Li *et al.*, 2018] Fengfu Li, Hong Qiao, and Bo Zhang. Discriminatively boosted image clustering with fully convo-

lutional auto-encoders. *Pattern Recognition*, 83:161–173, 2018.

[Li *et al.*, 2021] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8547–8555, May 2021.

[Lomov *et al.*, 2021a] Ildar Lomov, Mark Lyubimov, Ilya Makarov, and Leonid E Zhukov. Fault detection in tennessee eastman process with temporal deep learning models. *Journal of Industrial Information Integration*, 23:100216, 2021.

[Lomov *et al.*, 2021b] Ildar Lomov, Mark Lyubimov, Ilya Makarov, and Leonid E. Zhukov. Fault detection in tennessee eastman process with temporal deep learning models. *Journal of Industrial Information Integration*, 23:100216, 2021.

[Niu *et al.*, 2020] Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 735–751, Cham, 2020. Springer International Publishing.

[Pozdnyakov *et al.*, 2024] Vitaliy Pozdnyakov, Aleksandr Kovalenko, Ilya Makarov, Mikhail Drobyshevskiy, and Kirill Lukyanov. Adversarial attacks and defenses in automated control systems: A comprehensive benchmark. *arXiv preprint arXiv:2403.13502*, 2024.

[Reinartz *et al.*, 2021] Christopher Reinartz, Murat Kulahci, and Ole Ravn. An extended tennessee eastman simulation dataset for fault-detection and decision support systems. *Computers & Chemical Engineering*, 149:107281, 2021.

[Ricker and Lee, 1995] NL Ricker and JayHyung Lee. Nonlinear modeling and state estimation for the tennessee eastman challenge process. *Computers & chemical engineering*, 19(9):983–1005, 1995.

[Rieth *et al.*, 2018] Cory A. Rieth, Ben D. Amsel, Randy Tran, and Maia B. Cook. Issues and advances in anomaly detection evaluation for joint human-automated systems. In *Advances in Human Factors in Robots and Unmanned Systems*, pages 52–63. Springer International Publishing, 2018.

[Savchenko *et al.*, 2022] Andrey V Savchenko, Kirill V Demochkin, and Ivan S Grechikhin. Preference prediction based on a photo gallery analysis with scene recognition and object detection. *Pattern Recognition*, 121:108248, 2022.

[Tao *et al.*, 2020] Hongfeng Tao, Peng Wang, Yiyang Chen, Vladimir Stojanovic, and Huizhong Yang. An unsupervised fault diagnosis method for rolling bearing using stft and generative neural networks. *Journal of the Franklin Institute*, 357(11):7286–7307, 2020.

[Taqvi *et al.*, 2021] Syed Ali Ammar Taqvi, Haslinda Zabiri, Lemma Dendena Tufa, Fahim Uddin, Syeda Anmol Fatima, and Abdulhalim Shah Maulud. A review on data-driven learning approaches for fault detection and diagnosis in chemical processes. *ChemBioEng Reviews*, 8(3):239–259, 2021.

[Van Gansbeke *et al.*, 2020] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 268–285, Cham, 2020. Springer International Publishing.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Venkatasubramanian *et al.*, 2003] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, and Surya N. Kavuri. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311, 2003.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA, 20–22 Jun 2016. PMLR.

[Yang *et al.*, 2017] Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3861–3870. JMLR.org, 2017.

[Zhang and Qiu, 2022] Shuyuan Zhang and Tong Qiu. Semi-supervised lstm ladder autoencoder for chemical process fault diagnosis and localization. *Chemical Engineering Science*, 251:117467, 2022.

[Zheng and Zhao, 2020] Shaodong Zheng and Jinsong Zhao. A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis. *Computers & Chemical Engineering*, 135:106755, 2020.