UKAEA-RACE-PR(23)03

N. Petkov, H. Wu, R. Powell

Self-supervised deep representation learning adversarial autoencoder: Deep domain adaptation for data-based fault diagnostics.

Self-supervised deep representation learning adversarial autoencoder: Deep domain adaptation for data-based fault diagnostics.

N. Petkov, H. Wu, R. Powell

This is a preprint of a paper submitted for publication in
IEEE Transactions on Pattern Analysis and Machine Intelligence

Self-supervised deep representation learning adversarial autoencoder: Deep domain adaptation for data-based fault diagnostics

Nikola Petkov, Huapeng Wu, Roger Powell

July 13, 2022

Abstract

Calibration data for condition monitoring (CM), often originates from equipment running in different environment conditions, and operational settings. Due to the uneven distribution of the data, the performance of traditional machine learning approaches for CM can easily be skewed in favour of operating conditions with larger data distributions. This paper presents a novel unsupervised machine learning methodology for addressing the problem of domain adaptation for CM. A self-supervised deep representation learning adversarial autoencoder (SR-AAE) is proposed to model the latent space as a sum of two vectors: a categorical cluster identifier and a Gaussian distribution style vector. SR-AAE is regularised using a proposed self-supervision method which recycles the data samples back through the same network in order to strengthen the performance of the encoder model. Fault diagnostics is accomplished in two stages: domain adaptation by SR-AAE, and fault diagnostics by temporal variation shift monitoring of the flattened reconstruction error by principal component analysis (PCA). The proposed methodology is evaluated on FD004 turbofan engine degradation datasets from Commercial Modular Aero-Propulsion System Simulation (C-MAPSS). The results demonstrate that the proposed methodology is able to learn clear disentangled representations of the operating conditions in the latent space, and the approach shows excellent results in the task of domain adaptation for eliminating bias in the reconstruction error estimation. The results show 99.524% accuracy in binary classification of healthy/faulty state, without any prior knowledge of the faulty state. Our work represents a novel approach towards fault diagnostics in cases where only data from the healthy state is present.

Keywords: Fault diagnostics, Domain adaptation, Condition monitoring, Adversarial Autoencoder

1 Introduction

Lack of degradation data remains one of the fundamental challenges for development of data-based condition monitoring (CM) systems. Degradation data is often hard to find or produce due to the large costs involved and concerns about how well the artificial data represents real scenarios. On the other hand, there is usually a large amount of data from equipment operating in a nominal (healthy) manner. The healthy data commonly originates from equipment running in different operating and environment conditions with uneven data distributions. If the distribution of the data is not accounted for in the design of a CM system, the CM system can show a severely skewed performance across domains. The general assumption when using traditional machine learning approaches for CM is that both the train and test datasets have identical data distributions. As the operating and environment conditions change in time, it is important that the machine learning algorithm can perform with similar performance throughout the different domains of operation independently. Methods trying to solve this problem are referred to as domain adaptation methods [1].

Recently, there have been an increasing number of research works on feature disentanglement for domain adaptation using unsupervised machine learning approaches [2], [3], [4], [5], and some of these methods have achieved promising results for feature disentanglement in fault diagnostics and anomaly detection. In 2018, Akcay et al. [6] used generative adversarial networks (GAN) [7] for anomaly detection in X-ray images and achieved superior results to previous state-of-the-art approaches. In 2019, Zheng et al. [8] proposed a fault diagnostics approach using dual GAN architecture for generating artificial train samples. This method showed superior results in fault diagnostics for imbalanced data samples. In 2020, Li et al. [9] achieved superior results using a variation of an autoencoder (AE) neural network [10] for domain alignment of features with both inter-dimension and inter-sample correlations. One of the drawbacks of GAN-based fault diagnostics is, however, that due to the nature of samples generation from pure random noise, a decrease in performance can be manifested in terms of over-regularisation.

The aim of this study is to develop a novel and fully unsupervised machine learning methodology for fault diagnostics with advanced domain adaptation properties, which provides a clear separation of healthy and faulty data, works well with a limited amount of healthy data, and does not require any degradation data.

Our work is based on the unsupervised machine learning method known as an adversarial autoencoder (AAE) [11]. We propose an extension to this approach that is able to achieve self-supervision by recycling the reconstructed samples from the output of the AAE, and feeding them back in at the input of the AAE for increased neural network (NN) model performance. This method is called a self-supervised deep representation learning adversarial autoencoder (SR-AAE). A novel two stage fault diagnostics framework methodology is proposed where the SR-AAE is used for domain adaptation and principal component analysis (PCA) [12] is used for monitoring of the temporal variance shift of the SR-AAE

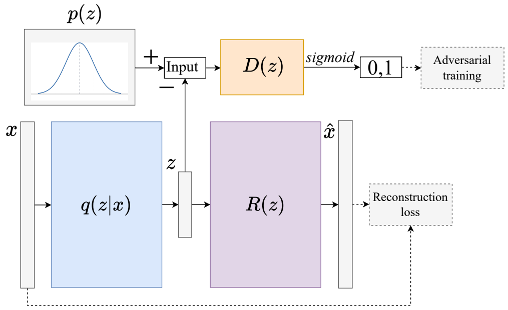

Figure 1: Basic structure of an adversarial autoencoder.

reconstruction error.

1.1 Adversarial Autoencoder

Recent advancements in unsupervised machine learning domain, represent potentially valuable tools for fault diagnostics with domain adaptation. In 2006, Hinton et al. [10] presented the autoencoder neural network. This architecture is mainly used for generating low level representations of highly non-linear datasets by using backpropagation and gradient descent, which is achieved by forcing the data through a bottleneck structure in the latent layer in the NN model. In terms of domain adaptation, the AE aligns the discrepancy between domains on the input dataset by minimizing the reconstruction error and learning invariant and transferable representation across domains in the latent space. In 2014, Goodfellow et al. [7] proposed a different novel framework for creating generative NN models called a generative adversarial network (GAN). The proposed framework used a novel adversarial training methodology to simultaneously train two distinct models, a generator G and a discriminator D. The models are trained by competing in a twoplayer minimax game described in Equation 1. The adversarial training methodology proposed in [7] can be applied across different domains and offers outstanding regularisation properties.

$$\min_G \max_D E_{x\sim p_{data}}[\log D(x)] + E_{z\sim p(z)}[\log(1 - D(G(z))] \tag{1}$$

In 2016, Goodfellow et al. [11] proposed an approach for a probabilistic autoencoder NN called an adversarial autoencoder (AAE) (Figure 1). AAE aims to model the latent space $z \sim q(z)$ (Equation 2) according to an imposed prior distribution $p(z)$ while jointly minimising the reconstruction error between the input dataset and target dataset. The modelling of the latent space to the imposed prior distribution is achieved using the adversarial training methodology described in (1). Using this approach, the AAE learns to convert the data distribution $p_d(x)$ to the imposed prior distribution $p(z)$ while learning a deep generative model that maps the learned prior distribution to the data distribution.

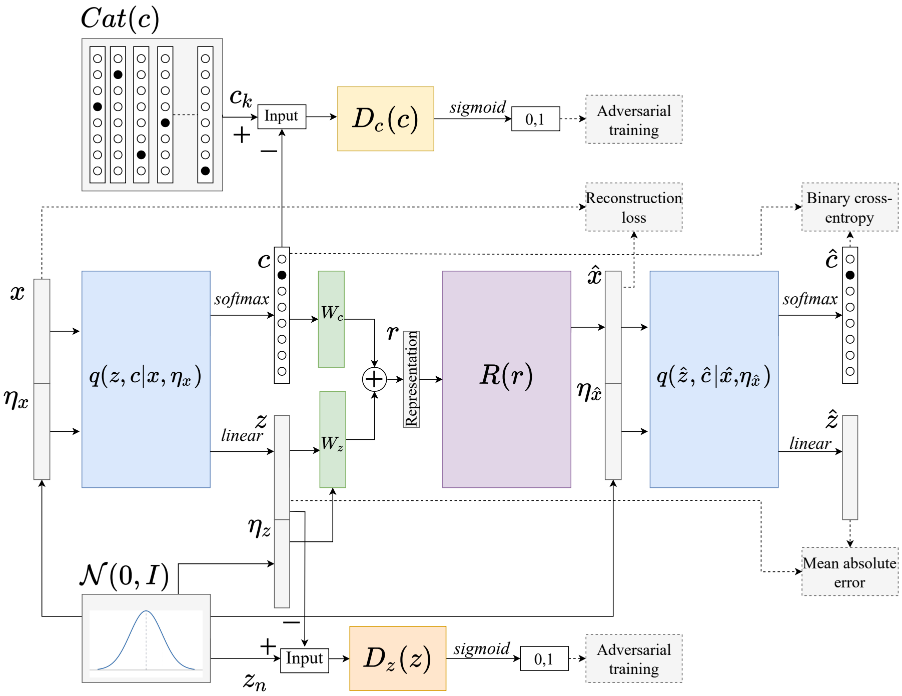

Figure 2: Self-supervised deep representation learning adversarial autoencoder (SR-AAE).

$$q(z) = \int_x q(z|x)p_d(x)dx \tag{2}$$

2 Self-supervised deep representation learning adversarial autoencoder (SR-AAE)

A more advanced AAE architecture is shown in Figure 2. The input of the network is represented as a function $f(x, \eta_x)$ where x is sampled from the data distribution p_d, and η_x is sampled from $\mathcal{N}(0, I)$. The network $q(c, z|x, \eta_x)$ represents the conditional probability of the latent space in terms of the input data distribution, and when optimised by a back propagation method, the network outputs the marginal probabilities $c, z \sim q(c, z)$ (3). A dual prior distribution that consists of both categorical and Gaussian distributions is imposed on the c and z latent space. Adversarial training loss, (7) and (8), is used to match the latent space to the prior distributions. The latent space z is then extended with an additional Gaussian noise vector η_z as a source of randomness. The extended Gaussian latent space vector z and the categorical latent space vector c are projected (embedded) to vectors of the same length using the matrices W_z and W_c, and they are combined using point-wise summation. This sum represents the latent data representation. The representation vector r is then fed to function $R(r)$ which reconstructs the input x as $\hat{x}$. The reconstruction error L_R (9) is backpropagated through the network to learn the updated weights. In order to achieve self-supervision, we join the estimated $\hat{x}$ by a Gaussian distribution noise vector $\eta_{\hat{x}}$ and use the encoder network q to generate $\hat{c}$ and $\hat{z}$. We then do additional back-propagation passes on the binary cross entropy loss between the initial vector c and the estimated vector $\hat{c}$ (10), and the mean absolute error between z and $\hat{z}$ (11).

$$c, z \sim q(c, z) = \int_x \int_{\eta_x} q(c, z|x, \eta_x) p_d(x) p_{\eta_x}(\eta_x) d\eta_x dx \tag{3}$$

$$r = \begin{bmatrix} z \\ \eta_z \end{bmatrix} W_z + cW_c \tag{4}$$

$$\hat{x} = R(r) \tag{5}$$

$$L = L_{D_c} + L_{D_z} + L_R + \lambda_1 L_{bce} + \lambda_2 L_{mae} \tag{6}$$

$$L_{D_c} = \min_Q \max_{D_c} E_{c_k \sim Cat(c)}[\log D_c(c_k)] + E_{x \sim p_d}[\log(1 - D_c(Q(x)))] \tag{7}$$

$$L_{D_z} = \min_Q \max_{D_z} E_{z_n \sim \mathcal{N}(0,I)}[\log D_z(z_n)] + E_{x \sim p_d}[\log(1 - D_z(Q(x)))] \tag{8}$$

$$L_R = \sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{n} \tag{9}$$

$$L_{bce} = \sum_{i=1}^{n} [c_i \log \hat{c}_i + (1 - c_i) \log(1 - \hat{c}_i)] \tag{10}$$

$$L_{mae} = \sum_{i=1}^{n} \frac{|z_i - \hat{z}_i|}{n} \tag{11}$$

$$\hat{c}, \hat{z} \sim q(\hat{c}, \hat{z}) = \int_{\hat{x}} \int_{\eta_{\hat{x}}} q(\hat{c}, \hat{z}|\hat{x}, \eta_{\hat{x}}) p_{AAE}(\hat{x}) p_{\eta_{\hat{x}}}(\eta_{\hat{x}}) d\eta_{\hat{x}} d\hat{x} \tag{12}$$

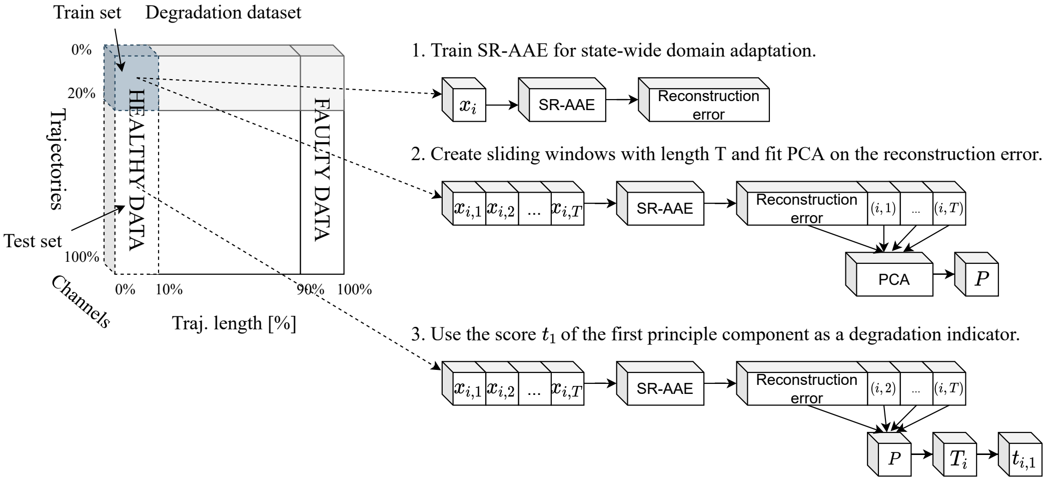
Train set
Degradation dataset
0%
20%
Trajectories
HEALTHY DATA
FAULTY DATA
Test set
100%
Channels
0%
10%
90%
100%
Traj. length [%]
1. Train SR-AAE for state-wide domain adaptation.
x_i
SR-AAE
Reconstruction error
2. Create sliding windows with length T and fit PCA on the reconstruction error.
$x_{i,1}$ $x_{i,2}$ ... $x_{i,T}$
SR-AAE
Reconstruction error
$(i,1)$ ... (i,T)
PCA
P
3. Use the score t_1 of the first principle component as a degradation indicator.
$x_{i,1}$ $x_{i,2}$ ... $x_{i,T}$
SR-AAE
Reconstruction error
$(i,2)$ ... (i,T)
P
T_i
$t_{i,1}$

Figure 3: Self-supervised deep representation learning adversarial autoencoder (SR-AAE) for fault diagnostics.

2.1 SR-AAE for fault diagnostics

A first step for data-based CM is identification of calibration data (healthy) and faulty data inside the data set (Figure 3), and separation of the dataset into train and test sets. For the purpose of our research, we assume that an initial 10% of the data trajectory length originates from equipment running in a healthy state. We assume that we have no knowledge of the equipment (no data) running outside healthy boundaries (initial 10%). The entire dataset is then normalised using the healthy train dataset parameters.

We use SR-AAE to align the variance of the reconstruction error R relative to the state domain on the input data distribution p_d. This alignment operation represents an intermediary step in our fault diagnostics approach, and removes the bias of the reconstruction error R for each operating condition. In order to capture the variance in the time domain of R, we need to form sliding windows of length T from the healthy train dataset $x_{T_i} = [x_{i,1}, x_{i,2}, ..., x_{i,T}]$, and align them using SR-AAE. PCA is a method for data analysis that tries to find a transformation matrix P (loadings matrix) that minimises the data variance from its axes (principal components) (13). Each column of the loadings matrix P is an eigenvector of $(R_T)^T R_T$ and represents a single principal component $\vec{p_i}$. The matrix T represents the projection (scores) of R_T to P. We fit the PCA transformation to the aligned reconstruction error windows $R_{T_i} = [R_{i,1}, R_{i,2}, ..., R_{i,T}]$ of the healthy train dataset. R_{T_i} represents the projection (score) of the R-windows to the axis of the PCA transformation that captures the most variance in the R-windows. We use $t_{i,1}$ (14) as a fault indicator in this research.

$$T = R_T P \tag{13}$$

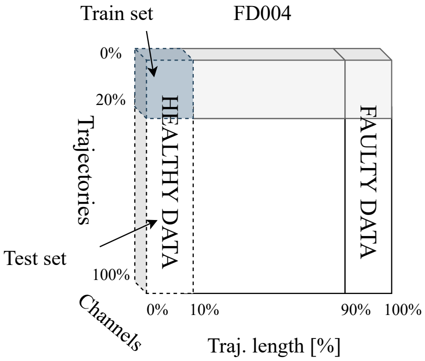

Figure 4: Anatomy of the FD004 dataset.

$$[t_{i,1}, t_{i,2}, ..., t_{i,T}] = R_{T_i}[\vec{p_1}; \vec{p_2}; ..., \vec{p_T}] \tag{14}$$

3 Fault diagnostics on a turbofan jet engine

The C-MAPSS [13] dataset represents simulated degradation data of 4 different flight modes of a turbofan jet engine that was created from NASA data for the 2008 Prognostics and Health Monitoring (PHM) competition. We evaluate our proposed methodology on the FD004 dataset from the C-MAPSS data. The FD004 dataset is a collection of 250 run-to-failure trajectories from a simulated turbofan jet engine that originate from 6 different operating conditions and results in two different failure modes. The data consists of 21 sensor measurements and 3 operational settings indicators (24 data channels). The lengths of the degradation trajectories are in ranges from 100 to 450 time steps. For the purpose of dividing the data into train and test datasets, the order of the trajectories is first randomised, 20% of the data trajectories are selected for training, and the remaining 80% of the trajectories are selected for testing (Figure 4). The initial 10% of the trajectory lengths is considered healthy and the last 10% of each of the trajectories is considered faulty. Our method only needs the healthy data for the purpose of fault diagnostics and does not require faulty data to identify the failure threshold. For the purpose of our research, some of the channels of data that do not show large variances are discarded, and the only channels used are: $2, 3, 4, 6, 7, 8, 11, 12, 13, 15, 16, 17, 18, 24$ and 25. The neural network shown in Figure (5) is a model of the proposed SR-AAE methodology.

It is trained only using the healthy train data (Figure 4) for 1000 epochs. Each training epoch takes approximately 2 seconds on a high-performance laptop. The size of c used in this case study is 12, which allows the NN to identify up to 12 clusters. The size of z is chosen as 2. The Gaussian noise vectors η all have the same size, that is, 7. The embedding size for the vector z concatenated

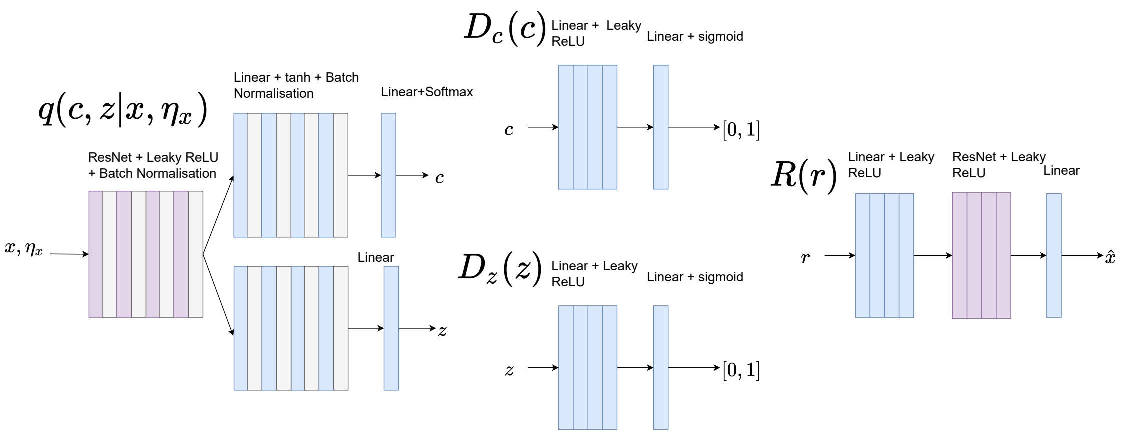

Figure 5: Architecture of the neural networks model used for SR-AAE. The size of each layer in the q and R models is 2048. The size of each layer in D_c and D_z is 512.

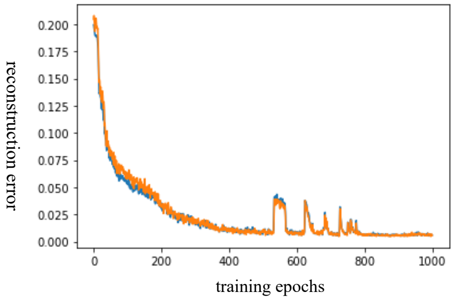

Figure 6: Trend line of the reconstruction loss on the train dataset (blue), and the reconstruction loss on the test dataset (orange) during 1000 epochs of training.

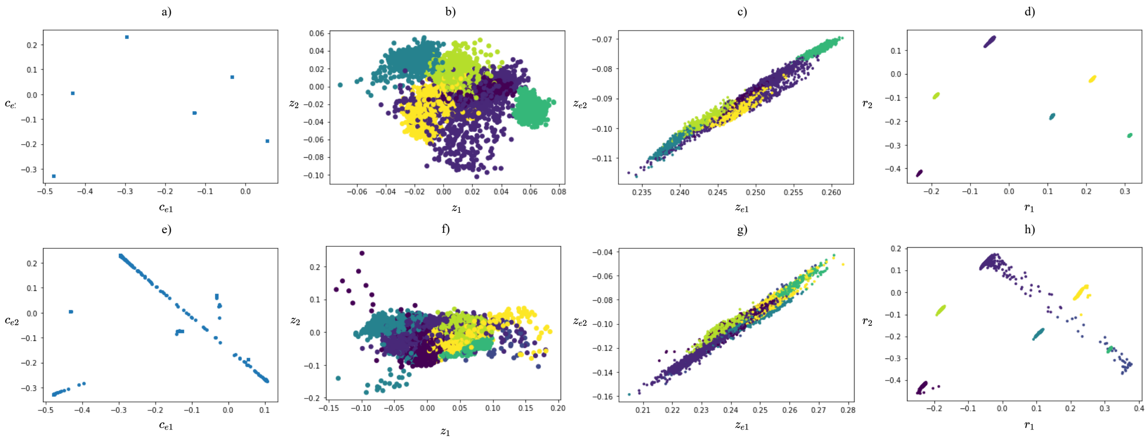

Figure 7: Visualisation of the latent space model on the validation dataset for the healthy and faulty data. The healthy data representations are shown in a), b), c), and d), where c_e is the embedding of c, z_e is the embedding of z, and r is the learned representation as a sum of both embeddings. The faulty data representations are shown in e), f), g) and h).

with the Gaussian noise vector η_z is 2, and is equal to the embedding size of the vector c in order to enable pointwise summation between both embeddings. We use ResNets [14] in the networks q and R for learning of deeper features which is manifested in faster convergence while training. We use dual heads for the network q: one for the output vector c and one for z. $tanh$ activation functions are used in the dual heads for added stability of the model training. The training process shows good convergence properties (Figure 6). It is important to note that the learning rate (LR) of the discriminator NN D_x and D_z should be an order of magnitude larger that the LR of the decoder network R, which should also be an order of a magnitude larger than the LR of the encoder NN q. This arrangement allows the network to first learn the adversarial training discriminators in order to aid the learning and domain adaptation of the decoder network R. The NN q is the most important NN because the other NN models depend on its output. Therefore, it has the lowest LR to allow the discriminators and the decoder NN to adapt to its outputs. A comparison of the manifolds in the latent space of the SR-AAE between the healthy data, and the faulty data in the test dataset is presented on Figure 7.

The results show 99.524% accuracy for classification tasks on the healthy test set using the latent vector c, represented as different colours in the data points of the plots in Figure 7. From the subplots d) and h) in Figure 7, we can observe that the proposed methodology works well for separating the outlier data far away from the healthy data, which is manifested as a separate cluster in h). Figure 8 a) and b) show the disentangled reconstruction error on the test dataset, and the density function on the healthy and faulty test data

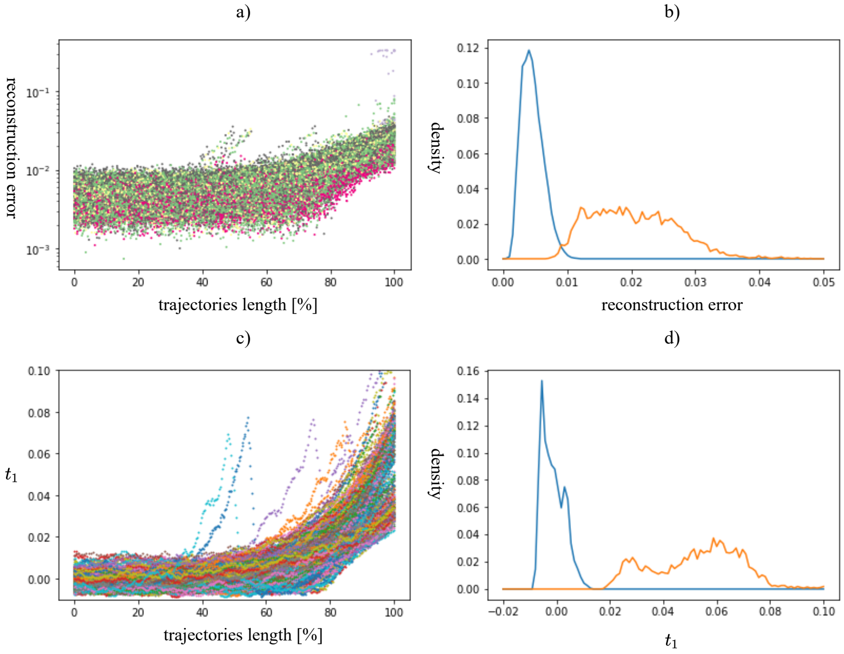

Figure 8: Visualisation of the domain invariant reconstruction error on the validation dataset a) and the PCA-filtered temporal reconstruction error c). The density functions of both approaches for the healthy and faulty validation datasets are shown in b) and d) respectively.

samples. Figure 8 c) and d) show the temporal PCA filtered reconstruction error with sliding window of size 12, and its density functions. The results show 0.00018 probability of overlap between the healthy and faulty data distributions of the reconstruction error, and 0 probability of overlap for the temporal PCA filtered reconstruction error. The performance of the temporal PCA filtered reconstruction error increases by increasing the sliding window length. The fault threshold is selected as the maximum value t_1 on the healthy train dataset (0.010858). As shown in Figure 9, our method is able to achieve 99.524 percent accuracy for classification of healthy/faulty states from the healthy and faulty test datasets.

4 Discussion

Current state-of-the-art methodology for fault diagnostics applied on the FD004 dataset [8] uses the modified infoGAN network called GAN for Failure Prediction (GAN-FP) for two class classification between healthy and faulty data samples separated in an 85% and 15% ratio respectively. GANFP and similar approaches are make an initial assumption that a sufficient amount of failure

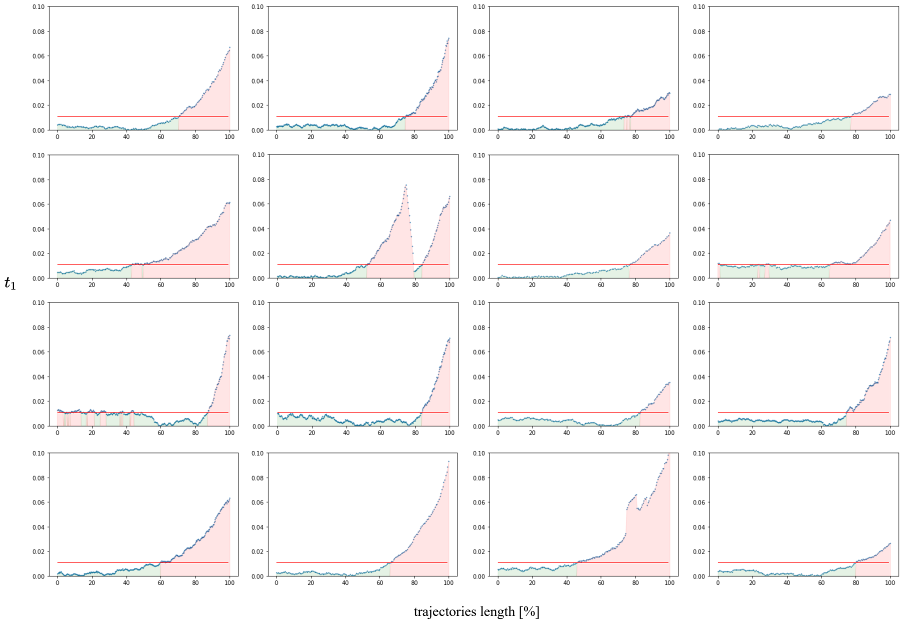

Figure 9: Visualisation of the performance of the proposed methodology for fault diagnostics on a random selection of 16 trajectories from the test set. The threshold line (red) is estimated as the maximum value (0.010858) of the proposed t_1 health indicator on the healthy train set.

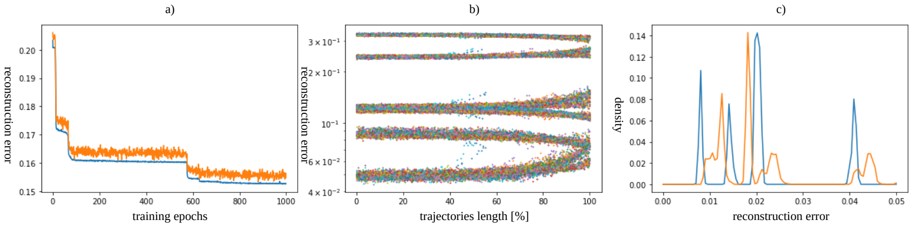

Figure 10: Application of a variational autoencoder (VAE) for fault diagnostics on the FD004 dataset without an initial domain adaptation.

data is available. In reality, access to failure data is one of the main impediments to effective NN development for fault diagnostics. Threshold-based approaches are suitable when failure data is not available. However, using threshold-based approaches for highly non-linear datasets is not a straightforward process. We solve this problem by using the proposed SR-AAE method for domain adaptation followed by a threshold-based approach on the flattened reconstruction error. The methodology proposed in our research is effectively able to put tight boundaries around the healthy data distribution and estimate any anomalies in the system with high accuracy, reporting for early signs of developing failure modes. Figure 10 presents a visualisation of the effect of data imbalance on reconstruction error when a plain variational autoencoder (VAE) is used. The bias in the reconstruction error for each operating condition is evident in Figure 10 b), and c) when compared to the proposed approach in Figure 8.

5 Conclusion

In this paper we introduced SR-AAE, a novel framework that utilises self-supervision and data recycling to increase the performance of a plain adversarial autoencoder (AAE) network and its application to fault diagnostics using temporal variance shift monitoring of the reconstruction error. We proposed a threshold-based two step approach for fault diagnostics where we use domain adaptation in the state space, and temporal variance shift monitoring in the time domain. Our results show that using this approach, we are able to achieve 99.524% classification accuracy on healthy against faulty state of the data. The proposed approach could potentially be applied for anomaly detection in other related domains where faulty data is scarce.

6 Acknowledgements

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 EUROfusion) and from the EPSRC [grant number EP/T012250/1]. To obtain further information on the data and models underlying this paper please contact PublicationsManager@ukaea.uk*. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

[1] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *CoRR*, vol. abs/2010.03978, 2020.

[2] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *CoRR*, vol. abs/1606.03657, 2016.

[4] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II1431II1439, JMLR.org, 2014.

[5] W. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *CoRR*, vol. abs/1709.07902, 2017.

[6] S. Akcay, A. A. Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," *CoRR*, vol. abs/1805.06725, 2018.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[8] S. Zheng, A. K. Farahat, and C. Gupta, "Generative adversarial networks for failure prediction," *CoRR*, vol. abs/1910.02034, 2019.

[9] H. Li, S. Wang, R. Wan, and A. C. Kot, "Gmfad: Towards generalized visual recognition via multilayer feature alignment and disentanglement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1289–1303, 2022.

[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015.

[12] I. Jolliffe, *Principal Component Analysis*, pp. 1094–1096. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[13] A. Saxena and K. Goebel, “Turbofan engine degradation simulation data set,” *NASA Ames Prognostics Data Repository*, 01 2008.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.