Domain Adaptation for Fault Detection in Civil Nuclear Plants

Henry Wood[1], Felipe Montana [2], Visakan Kadirkamanathan [3], Andy Mills [4], Will Jacobs[5],

[1, 2, 3, 4] *The Department of Automatic Control and Systems Engineering*
The University of Sheffield
Amy Johnson Building, Portobello Street, Sheffield, S1 3JD, United Kingdom
henry.wood@sheffield.ac.uk
f.montana-gonzalez@sheffield.ac.uk
visakan@sheffield.ac.uk
a.r.mills@sheffield.ac.uk
w.jacobs@sheffield.ac.uk

ABSTRACT

Recent domain adaptation approaches have been shown to generalise well between distant data domains achieving high performance in machine fault detection through time series classification. An interesting aspect of this transfer-learning inspired approach, is that the algorithm need not be exposed to fault data from the target domain during training. This promotes the application of these methods to environments in which fault data is unfeasible to obtain, such as the detection of loss-of-coolant accidents (LOCA) in nuclear power plants (NPPs).

A LOCA is a failure mode of a nuclear reactor in which coolant is lost due to a physical break in the primary coolant circuit. If undetected, or not managed effectively, a LOCA can result in reactor core damage.

Three high-fidelity physics based models were created with divergent behaviour that represent different data domains. The first model is used to generate source domain data by simulating labelled training data under both nominal and LOCA conditions. The second and third models act as surrogates of real plants and are used to generate target domain data, i.e. to simulate nominal data for training and LOCA condition data for validation.

Several deep-learning feature encoders (with varying levels of connectivity) were applied to this LOCA detection problem. Among these, a 'Baseline' encoder was used to quantify the improvement that domain adaptation techniques make to LOCA detection performance under large domain divergences.

Classification accuracy for each model is explored within the context of LOCA break size and location within each plant model.

The proposed method for LOCA detection demonstrates how the dependence upon sparse accident-specific data can be alleviated through the use of domain adaptation. Detection capability of the LOCA condition is maintained even when no data examples are available in the target domain.

1. INTRODUCTION

There is an opportunity in the nuclear industry to adopt data-driven methods to help maintain the safe operation of critical systems, both as a result of the improved availability of sensing instrumentation and the rapid evolution of network architectures for fault detection (Gomez-Fernandez et al., 2020). A plethora of approaches for identifying abnormal transient behaviour, such as a LOCA, exist with foundations in probabilistic methods (Aldemir, 2013), complex fluid-structure interaction models (Mahmoodi et al., 2011) and Markov modelling (Sakurahara et al., 2019). Although effective in finite environs, conventional methods suffer when attempting to compensate for the lack of available labelled fault data in the nuclear domain.

A LOCA occurs when a physical break in the reactor coolant system releases coolant faster than recovery systems can replenish it. This increases the temperature of the core, which can damage the plant and potentially release reactivity. Detection of this transient behaviour is paramount to the safe operation of pressurised water reactors (PWRs).

Time series classification through deep learning methods has seen increased attention (Ismail Fawaz et al., 2019) in previous years, with myriad techniques being derived to tackle a spectrum of fault detection problems (Wei & Keogh, 2006). The nuclear industry has received its share of attention in this regard, with neural network based methodologies tasked with

aiding the monitoring of numerous aspects of NPPs, including diagnosing the source of abnormalities in operation data (Lee et al., 2021) and tuning a digital twin to provide supplementary NPP data (Wang et al., 2021).

Existing examples of these methods simultaneously identify and characterise transients whilst making remaining useful life predictions (Rivas et al., 2024). Typically, though, these approaches rely upon the assumptions that similar quantities of nominal and faulty data exist, and that data gathered from differing NPP sources (physical plants and simulations alike) will share a similar data distribution.

One branch of deep-learning research aims at tackling this manner of problem through Transfer Learning. Current works in industrial contexts display impressive results regarding fault diagnosis with minimal labelled training data under diverse application domains (Y. Zhang et al., 2023), as well as combinations of global and local models providing more robust remaining useful life predictions (J. Zhang et al., 2023). Domain adaptation (a subset of Transfer learning where data sources share the same input space) can simultaneously make data gathered from multiple sources appear more similar, whilst separating sub-classes within those sources, eg. 'Normal' and 'Faulty' data (Qian et al., 2023).

Existing LOCA detection procedures that attempt to overcome the issue of the lack of available NPP accident data perform well in a limited range of operating conditions (Farber & Cole, 2020). The generalised knowledge available through leveraging transfer-learning from attainable model data has not yet been fully exploited in the context of LOCA detection. Domain adaptation allows the transfer of the knowledge contained in such a classifier on to a new domain containing previously unseen behaviour.

In this work, we introduce adaptations to current transfer learning based fault detection methods with application to the detection of the LOCA condition. The model design process was guided by system experts in order to construct data 'features' that well represent the NPP behaviour in both nominal and LOCA conditions. Results show how domain adaptation is able to retain the fault detection performance that is achievable for the labelled training data when it is applied to the surrogate models data domain.

2. Problem formulation

2.1. Domain adaptation overview

Consider data sampled from two distinct domains, Source (S) and Target (T). The data from each domain (x_S and x_T, respectively) posses different distributions. Additionally, suppose that class labels for the data sampled from the Target domain, y_T, are unavailable. Given the data is drawn from disparate distributions, conventional supervised methods cannot infer knowledge about the Target domain using data from the Source domain.

Domain adaptation provides methods for prediction of target domain labels y_T from target domain data x_T using the information present in source domain data and labels x_S and y_S. In this work, we will describe a feature extractor as an encoder: a network designed to construct a feature space Z using the distributions of x_S and x_T. The aim of the encoder is to provide a transformation through which the distributions of x_S and x_T appear similar to each other in the feature space Z.

The generalised feature space Z is used to aid classification for samples from the target domain, since the encoded representations of x_S and x_T are similar, and we have access to class labels for the source domain data, y_S. There exist many well documented methods by which the encoder can construct Z, with two of the most commonly used domain-adaptation specific measures being Mutual Information (MI) and Maximum Mean Discrepancy (MMD).

2.1.1. Mutual Information

MI is a statistical quantity that describes how much information one variable conveys about another. If we consider these variables in terms of the feature space representations of the input domain data, i.e: z_S and z_T (obtained from passing x_S and x_T respectively into the transformative encoder), then maximising the MI between the Target domain feature space representation (z_T) and the entire feature space (Z) will encourage the encoder to generate features that are generalised between the two input domains.

The MI between these specific variables can be expressed as a linear combination of the Shannon Entropy of each feature space representation (Chen et al., 2021). The Shannon Entropy for a distribution A is given by

$$H(A) = -\sum_{a \in A} P(a) \ln P(a). \tag{1}$$

If we state that Z_S and Z_T are the distributions of the feature space representations z_S and z_T respectively, then the MI becomes

$$\text{MI}(Z_T; Z) = -\sum_{z_S \in Z_S} P(z_S) \ln P(z_S) - \sum_{z_T \in Z_T} P(z_T) \ln P(z_T). \tag{2}$$

Maximising this quantity during training promotes the generation of features that convey the largest amount of shared information between the Target domain samples and the entire set of observed samples from each domain.

2.1.2. Maximum Mean Discrepancy

A brief description of the intended function of the MMD term will be sufficient for understanding its relevance to this work. The key principal that underpins MMD metrics is the idea that if two distributions are equal, then their statistical properties should also be equal. By using MMD, it is possible to perform a hypothesis test upon the functions that transform the input domain distributions into their encoded feature representations. These functions are embedded as a Hilbert space, a convenient mathematical construct which allows linear algebra to be applied to infinite-dimensional vectors.

Formally, the MMD between two distributions A and B on the sets X and Y can be calculated as

$$\begin{aligned}\text{MMD}(A,B) &= ||\mathbb{E}_{X\sim A}[\phi(X)] - \mathbb{E}_{Y\sim B}[\phi(Y)]||_H \\ &= \sup_{f\in H}(\mathbb{E}_{X\sim A}[f(X)] - \mathbb{E}_{Y\sim B}[f(Y)]), \quad (3)\end{aligned}$$

where f is a function in the Hilbert space H and ϕ is the transformation from the input set to the Hilbert space. The supremum means this is equivalent to taking the maximum of the mean difference between the distributions A and B.

In practice, the mean of the feature-space distributions is not known, so the MMD between the two feature-space distributions must be empirically estimated by

$$\begin{aligned}\text{MMD}(Z_S,Z_T) = &\frac{1}{m(m-1)}\sum_{i}^{m}\sum_{j\neq i}^{m}\phi(z_{S_i},z_{S_j}) \\ &- 2\frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n}\phi(z_{S_i},z_{T_j}) \\ &+ \frac{1}{n(n-1)}\sum_{i}^{n}\sum_{j\neq i}^{n}\phi(z_{T_i},z_{T_j}), \quad (4)\end{aligned}$$

where, m and n are the number of samples drawn from Z_S and Z_T, and ϕ is a Gaussian kernel representing the feature mapping transformation. A minimisation of MMD ensures that the distributions Z_S and Z_T are similar across each statistical moment, which aids in making predictions about the unlabelled Target domain.

2.1.3. Domain adaptation-oriented loss function

The Negative Log Likelihood, NLL, cost is used to penalise incorrect classification predictions made by the model and is given by

$$\text{NLL}(\theta) = -\sum_{i=1}^{k}\left(y_i\ln(\hat{y}_{\theta i}) + (1-y_i)\ln(1-\hat{y}_{\theta i})\right), \quad (5)$$

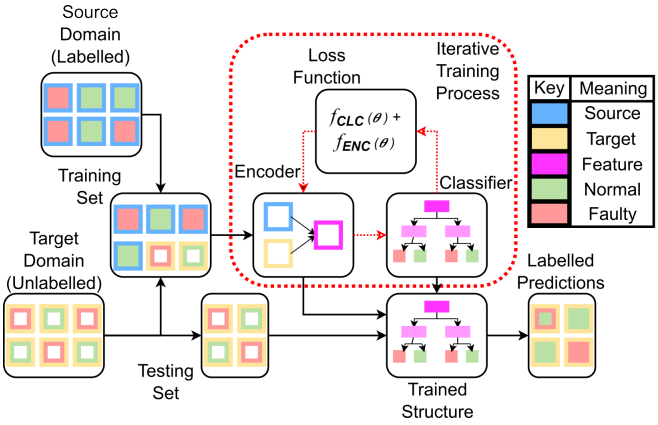

Figure 1. An example of an unsupervised domain adaptation approach. An encoder and classifier are trained simultaneously to generate both a representative feature space and accurate class predictions.

where, θ is a set of probabilities attributed to each class prediction, k is the number of predictions made, y is the true class of each sample and $\hat{y}$ is the predicted class.

To perform domain adaptation the following loss function is used:

$$\mathcal{L}_{DA} = \text{NLL}(\theta) + \text{MMD}(Z_S,Z_T) - \text{MI}(Z_T,Z). \quad (6)$$

This is a common form for a loss function seen in an unsupervised domain adaptation setting, visualised in Figure 1.

For comparison, this work will also use a simplified version of this loss function, $\mathcal{L}_S = \text{NLL}(\theta)$, to represent a loss function used by a conventional supervised learning approach.

3. Methodology

3.1. Data generation

In light of the lack of labelled relevant NPP accident data, RELAP5, a nuclear reactor modelling and simulation tool, was used to generate the data used in this work. Rather than perform domain adaptation between model generated data and real data collected from a plant, a high-fidelity physics models is used under three different configurations that represent data domains with varying levels of divergence between them. The model configurations represent 1) A large four-loop 3600MW civil nuclear plant with nominal historical usage, 2) A similar large plant with a greater historical power usage and 3) A small two-loop 50MW with nominal historical usage.

3.1.1. Model modifications

To replicate the full range of operating conditions of a civil nuclear plant, and for a machine learning framework in this

setting to be trained robustly, data containing examples of dynamic events are provided. These events come not only from the presence of LOCA/faults, but also reflect dynamism in the normal operation of a plant, such as reactivity insertion.

The specification of a general table used to define the core reactivity or power (depending upon the reactor kinetics model used by the script) proved sufficient for providing the kind of input-derived transient events required. Typically, these input demands have magnitude between 1-10% of the input reactivity/steam off-take of that observed at the reactor's steady state rated power output level.

A small degree of Gaussian noise was added to the input reactivity demand profile to simulate process noise. The scale of the input noise was less than 10% of the magnitude of the changes in input demands. Set-points and thresholds that define variable and logical trips for control systems in the model were perturbed to emulate differing operator characteristics on each run.

Each simulation was performed with or without the presence of a LOCA (hence being classed as 'Normal' or 'Faulty').

Breaks were inserted into the primary circuit of the reactor coolant system to simulate a LOCA. Breaks are simulated at the inlet and outlet of the hot and cold legs of the primary circuit, as well as at the outlet of the steam generator in the secondary circuit. All breaks used the counter-current flow model, with standard choking flow. The full abrupt change model was used meaning that all breaks occurred instantaneously, rather than develop throughout the course of one sample of time-series data. The breaks are modelled as a valve with given cross-sectional area. The cross-sectional area of the break-valve is adjusted to define the size of the break relative to the cross-sectional area of the pipe to which the break-valve is located. The break sizes are uniformly sampled in the range $[0.02\%, 0.2\%]$ of the area of the pipe for the 3600MW plant, and the range $[0.1\%, 1\%]$ of the area of the pipe for the small 50MW plant, representing very small breaks. Each simulation is run for 1000 seconds.

A summary of the numerical changes to the high-fidelity physics models is as follows:

- Transient operating power provided by control of reactor rod position or steam off-take at the steam generator. Operating power level varied between +-10% of the rated capacity of each plant.
- Gaussian process noise inserted with transient input signals, scaled to +- 1% of the rated capacity of each plant.
- Control system thresholds shifted by -2%, +2% or unchanged for each simulation.
- Breaks inserted with magnitudes in the range $[0.02\%, 0.2\%]$ and $[0.1\%, 1\%]$ of the cross-sectional area of the pipe in which they are located for the 3600MW and 50MW plant respectively.
- Each simulation is run for 1000 seconds.

3.1.2. Differences between models (domain divergence)

This work focuses on exploring the implication of an increasing divergence between data domains. In this context, this requires multiple high fidelity physics models from which to gather data. The first of the template models used describes a large Four-loop 3600MW PWR with characteristics designed to be a 'fictitious approximation' of values present in a Westinghouse plant.

To provide an example of a relatively small domain divergence, the 3600MW PWR model is used to provide data representative of the same plant at different stages in its operating cycle. To achieve this, different model initialisation applied that define different average operating power output for the first year of operation. An 'Underworked' version of this Four-loop plant was defined to have operated at 2400MW (significantly less than the 3600MW rated capacity) for its first year of operation. This model was used as the 'Source' domain model. A 'Typically worked' version of the same plant is defined to have operated at its rated capacity of 3600MW for the first year of it's operation. Additionally, pump speeds throughout the primary and secondary circuits are increased, allowing for different dynamics to manifest in this version of the plant. The model generated data used for the 'Target 1' domain.

A second, smaller Two-loop 50MW PWR plant was chosen to represent a more drastic domain change. This plant is simulated with the same relative changes in input reactivity and trip set-point attitudes, but possessing disparate dynamics and steady state behaviour owing to the different physical properties of the smaller plant. Data generated from this model is used for the 'Target 2' domain.

A quick understanding of the degrees of divergence between these domains can be gained by observing the first principal component of a principal component analysis performed on each domain, shown in Figure 2. The distribution of the first principal component differs greatly depending upon which domain the data comes from, although the difference between the Source and Target 2 domains is much larger than that between the Source and Target 1 domains. This domain divergence would cause a conventional supervised learning approach to suffer, due to the difficulty in transferring knowledge between disparate domains. Additionally, note the large degree of overlap that exists between 'Normal' and 'Faulty' classes of data in each domain. This implies that, since principal components analysis is a linear technique, a linear classifier would struggle to separate samples of data from different classes.

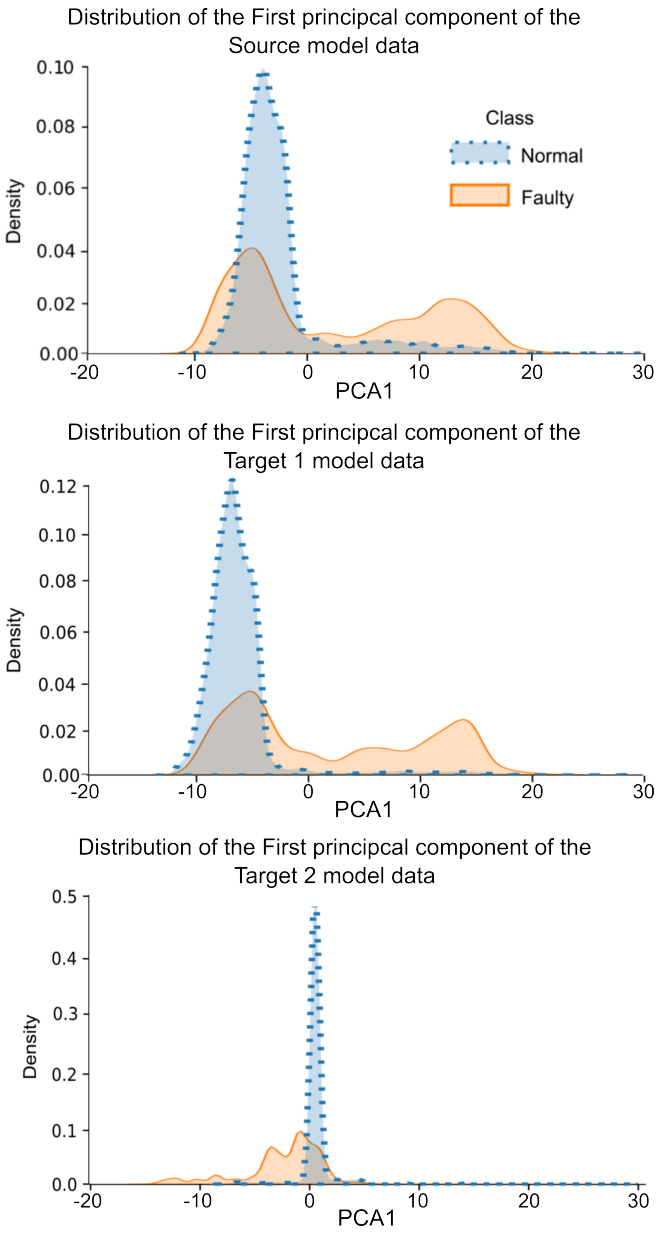

Figure 2. The distributions of the first principal components of the data from each of the three domains. Note the larger difference in distributions between the Source and Target 2 domains.

3.1.3. Processed data structure

Observations of the plants were made by simulating sensors in locations throughout both the primary and secondary circuits, listed in Table 1. In total, 15 data streams were extracted from each simulation, representing pressures, temperatures, mass-flow rates and valve states, with locations shown in Figure 3.

In addition to these measured values, the input reactivity profile (the power demand) and the reactor output power were supplied as part of the data vector provided to the domain adaptation network. Each batch of data the encoder receives is made of samples derived from both the Source and Target domains. At the encoder, these samples are unlabelled. Each sample x is an $N \times T$ vector representing T time-steps of N sensor readings. The input dimension of the encoder is $B \times N \times T$, where B is the number of samples used per batch.

Table 1. A list of the measured values used in this work.

Value specification		
ID	***Description***	***Units***
0	Cold-leg Coolant Pressure	P
1	Cold-leg Inlet Coolant Temperature	°C
2	Cold-leg Outlet Coolant Temperature	°C
3	Hot-leg Coolant Pressure	P
4	Hot-leg Inlet Coolant Temperature	°C
5	Hot-leg Outlet Coolant Temperature	°C
6	Pressuriser Relief Valve State	-
7	Main Steam Isolation Valve State	-
8	SG Feedwater Regulating Valve State	-
9	Cold-leg Coolant Mass Flow-rate	kg/s
10	Hot-leg Coolant Mass Flow-rate	kg/s
11	Reactor Coolant Pump Mass Flow-rate	kg/s
12	SG Feedwater Inlet Mass Flow-rate	kg/s
13	Pressurizer Inlet Mass Flow-rate	kg/s
14	Charging Pump Mass Flow-rate	kg/s

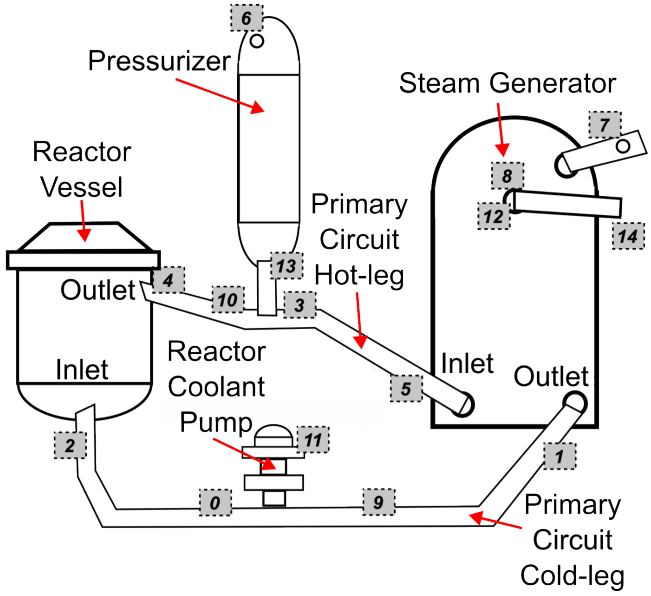

Figure 3. A simplified view of the primary coolant circuit of a PWR. The measured sensor values used in this work have their sensed locations numbered.

3.2. Model architecture

The network described in this work consists of an encoder, tasked with extracting generalised feature maps from the raw input vector data, and a classifier aimed at detecting the presence of 'Faulty' data. Different model architectures were were trialed and they are described in detail below.

3.2.1. Encoder variations

To investigate the impact that the internal structure of the feature-generation stage of the network has in the context of NPP time-series data, several forms of encoder were considered, with increasing degrees of 'connectivity' between different input sensor readings, visualised in Figure 4.

Baseline Encoder: Separate 1D convolutions

Two kernels per input sensor channel, with kernel grouping number set equal to the input dimension at each point in the encoder. This has the effect of training kernels without combining information from multiple sensor channels simultaneously. Kernels act along the time dimension of each sensor reading.

Aggregate Encoder: Summed 1D convolutions

Grouping number for convolutional layers set to 1, meaning kernels are passed along a single time-series sensor reading, before being combined in a weighted sum to generate a convolution which contains information from each sensor measurement simultaneously.

Recurrent, Fully-Connected Encoder: Gated Dense 1D convolutions

The summed 1D convolutions have been performed as above followed by a fully-connected layers Additionally, a gated recurrent unit (GRU) layer is used before the second set of convolutions.

These modifications attempt to allow the encoder to efficiently consider long-term dynamics that may be important to fault detection in this context.

3.2.2. Classifier & loss function

The classifier is shared by each of the different encoder variations described above. The classifier consisted of a series of fully connected layers followed by batch normalisation layers, shown in Figure 5. A dropout layer is included to encourage regularisation and avoid over-fitting.

4. Experimental validation

The loss function used by each model varies depending upon whether domain adaptation is used. When a model is applied without domain adaptation, the loss $\mathcal{L}_S$ is used. The loss $\mathcal{L}_{DA}$ is used when domain adaptation is required. The 'Baseline' model was tested in each data domain twice, once using $\mathcal{L}_S$ and once using $\mathcal{L}_{DA}$. The Aggregate Encoder and Recurrent Fully-Connected Encoder were tested on all data domains using the loss $\mathcal{L}_{DA}$.

4.1. Hyper-parameter tuning result

The model hyper-parameters were tuned heuristically for each model variant. Hyper-parameters were chosen as

Baseline Model:

- Learning rate: $8e-4$

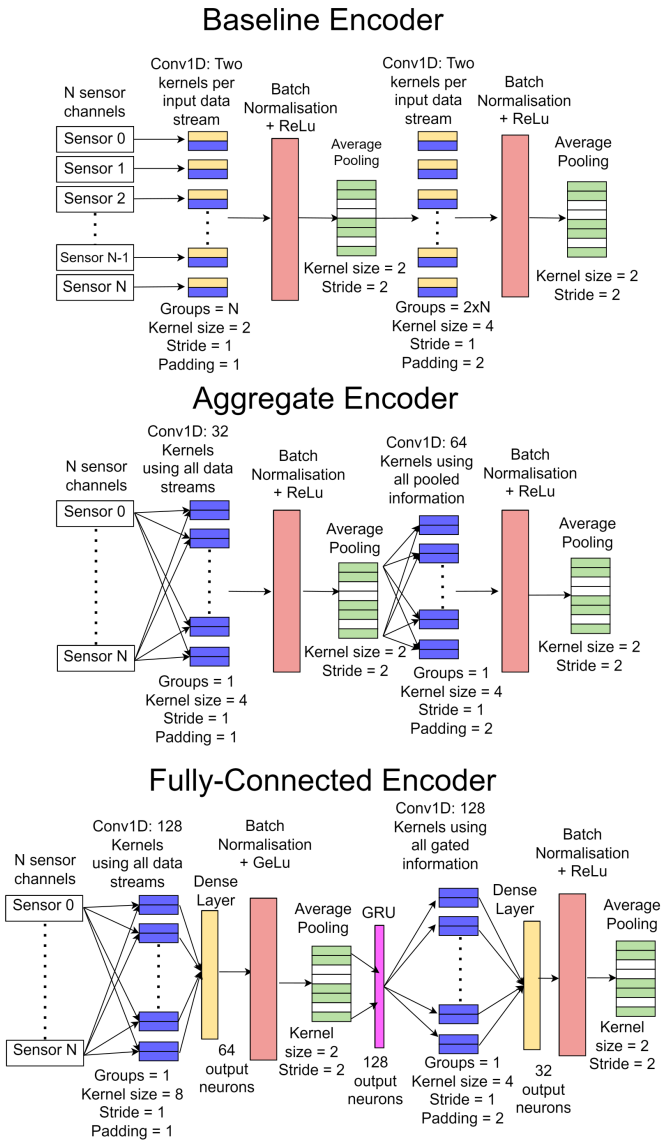

Figure 4. Architectures of three different encoders used in this work, with increasing degrees of connectivity between the input sensor readings.

- Dropout percentage for classifier input: $7.0\ \%$
- Scaling factor for MMD term in loss function: 0.5

Aggregate Encoder Model:

- Learning rate: $9e-4$
- Dropout percentage for classifier input: $4.1\ \%$
- Scaling factor for MMD term in loss function: 0.5

Fully-Connected Encoder Model:

- Learning rate: $6e-4$
- Dropout percentage for classifier input: $5.6\ \%$
- Scaling factor for MMD term in loss function: 0.5

4.2. LOCA detection

This section, detailing the results from this work, is divided into three parts covering, binary LOCA classification performance, detection performance by break size and detection performance by break location.

4.2.1. Binary classification performance

When tested on the the Source domain, each model performed similarly in classification accuracy, with over 93% of the samples observed being correctly classified as either 'Normal' or 'Faulty' for each model variant, as shown in Table 2. Without the necessity for domain adaptation in this case, each model was able to create a robust feature map of the distribution of data observed in the Source domain. The increased complexity of the Aggregate and Fully-Connected Encoders offered little to no benefit in this conventional supervised setting, with the training and testing sets being drawn from the same domain.

Disparities in model performance start to appear when the testing set is drawn from the Target 1 data domain. This testing set represents a slight shift in data distribution from the Source domain training set, which reveals the importance of the inclusion of model architectures specifically designed to aid in domain adaptation. The models utilising the more sophisticated domain adaptation-oriented loss function retain the majority of their classification performance when compared to test results from the Source domain, whilst the Baseline supervised model, using $\mathcal{L}_S$, suffers a sizeable reduction in classification accuracy.

This difference in performance is owed to the fact that the domain adaptation-oriented loss function contains terms that inform the encoder about the statistical properties of the generalised feature space that it is tasked with creating. Consideration of the MI between the entire feature space and the encoded representation of the Target 1 domain data helps to reduce the likelihood of encountering unlabelled Target 1 domain samples during testing that do not possess some information that the model has previously observed during training. Prompting this overlap in shared information increases that chances that the model will hold some 'relevant' feature space representation for these 'unique' unseen phenomena, which is crucial due to the high variability of the generated data. Additionally, minimisation of the MMD at the encoder aids the classifier with inferring class labels belonging to the unlabelled Target 1 domain data. This is explained by the fact that a reduction in MMD between the encoded representations of the Source and Target 1 domains implies that samples belonging to one class (for example, 'Faulty') that exist in one region within the encoded Source domain feature space, should exist in a similar 'relative' location within the encoded Target domain feature space. It is through this knowledge transfer that Source domain information can be leveraged to support Target domain class predictions.

The final domain shift, between training on Source domain model data and testing on the Target 2 domain data represents a more severe divergence between domains. As is evident from the average test accuracies, the conventional supervised model fails to bridge this gap between differing data distributions and records a poor performance of less than 50% classification accuracy. It is at this magnitude of domain divergence that the increased complexity of the Aggregate and Fully-

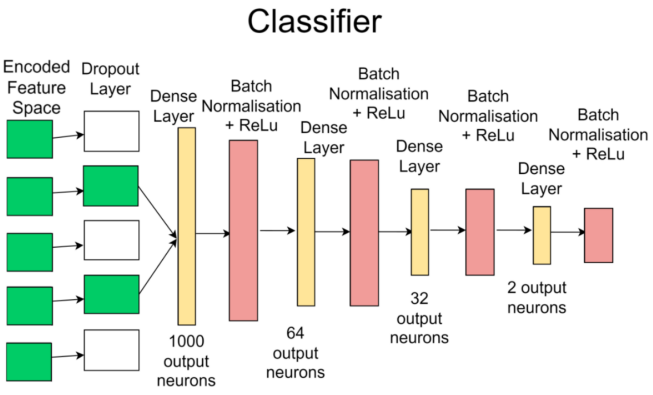

Figure 5. Architecture of the classifier used in this work.

Table 2. Average test accuracies (%) of each model type on each domain set, after being trained on the Source data domain.

Model Type	Source Domain	Target 1 Domain	Target 2 Domain
Baseline Supervised (No DA)	93.41	78.02	45.57
Baseline With DA	-	91.24	81.46
Aggregate Encoder	93.96	91.02	89.85
Fully-Connected Encoder	93.22	91.27	89.57

Connected Encoders demonstrate value. Although the DA-focused loss function is enough to restore the majority of the lost performance for the baseline model, the less-connected Encoder lacks the ability to consider the more disparate dynamics evident in the Target 2 domain.

In the context of this work, allowing the encoder to combine input sensor channels at the first convolutional layer before performing the convolution (such as in the Aggregate and Fully-Connected encoders), may allow the models to exploit class-specific relationships between sensed quantities. For example, in the presence of a LOCA, a brief divergence of primary circuit hot-leg temperature and pressure may occur. If these sensed values were provided to the same convolutional kernel at the input layer of the encoder, then the kernel could utilize this relative disparity to generate a recognisable identifier of a LOCA class sample. This improvement in performance compared to the 'Baseline With DA' model suggests that there is more information available with respect to the problem of LOCA detection if the sensed values are processed relative to each other, rather than processed in parallel.

Although the nature of LOCA simulated in this work vary drastically in magnitude and location within the modelled plants, there exists the possibility for other fault cases to transpire in an NPP. Without explicit knowledge of the existence of these faults (other than LOCA), the accurate classification of these samples as 'Faulty' would depend upon the similarity of these encoded samples to the 'Normal' encoded data. The performance of the models in this work on LOCA-specific fault detection is good, meaning the classifier used can differentiate between 'Normal' data, and all other LOCA data. If another fault case, previously unseen by the models described, manifested in a similar fashion to a LOCA, it is likely it would be identified as 'Faulty'. However, as can be observed in later analysis on model performance by break location, there can be large difference in classification accuracy for a single model across faults from multiple locations, so it would not be reliable to depend upon these methods as part of generic 'anomaly detection' techniques.

An understanding of the impact of the encoder in this work can be understood if the distributions of the encoded data from each domain are observed, shown in Figure 6. The data used in this figure are drawn from the Fully-Connected encoder. Viewing the first principal component of the post-encoder data from each domain reveals that the three domains appear much more similar to each other once expressed in the generalised feature space the encoder provides. As observed previously, the encoded distributions still share a large degree of overlap between 'Normal' and 'Faulty' data. Since principal components analysis is a linear transformation, this suggests that a linear classifier would struggle to reliably predict the class of unlabelled samples, and that the models used in this work which perform well must consider nonlinearities in the data.

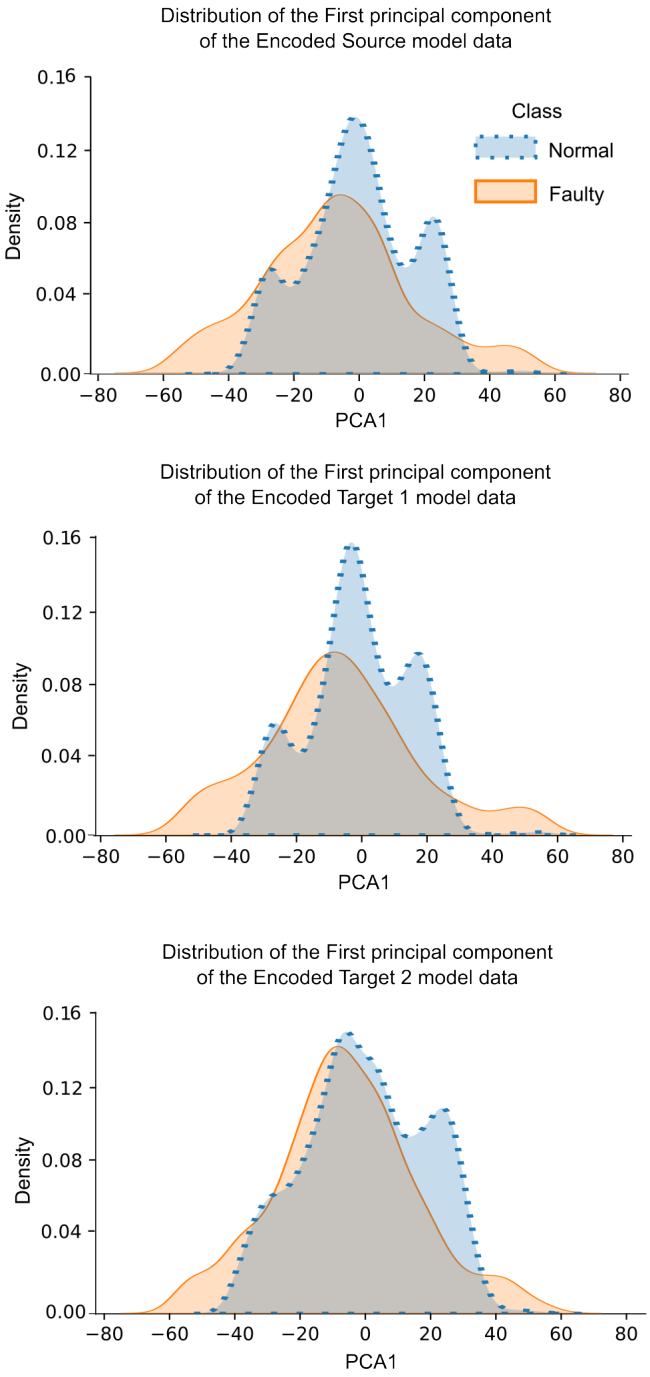

Figure 6. The distributions of the first principal components of the encoded data from each of the three domains. Note the more similar distributions compared to the first principal component of the input space.

4.2.2. LOCA detection by break size

An interesting perspective by which to consider the performance of these models in the context of NPP fault detec-

tion is to observe only the primary circuit breaks and identify the thresholds above which each model can always identify a LOCA. LOCA detection by break size results are shown in Figure 7.

In the Source domain setting, each model could reliably categorise fault data with break size above 0.1% of the pipe cross-sectional area as 'Faulty'.

The performance is retained when the models are tested on the Target 1 domain set, however the rate at which the baseline model without domain adaptation can successfully categorise samples below 0.1% is substantially lower than in the source domain.

The performance loss is exaggerated as the gap between domains increases further still: without considering domain adaptation, there is no size of primary circuit break that the baseline model can always categorize correctly as 'Faulty'. With the only alteration being the inclusion of domain-adaptation focused terms in the loss function, the Baseline model (With DA) can, on average, identify 20% more faults successfully.

The other models retain a substantial proportion of their ability to successfully categorize all break sizes, even in this most extreme domain divergence example.

4.2.3. LOCA detection by break location

The 'Faulty' samples in this work were not only drawn from a range of possible sizes, but also placed at varied locations throughout the primary and secondary circuit of each PWR. Primary circuit breaks occur at either the inlet or outlet of the hot or cold legs. Secondary circuit breaks were inserted at the outlet of the steam generator for the respective loop. LOCA detection by break location results are shown in Figure 8.

When tested in the native Source domain setting, both the Baseline Supervised and Aggregate Encoder models correctly classify all nominal operation data, along with a consistently high successful classification rate of primary circuit breaks as 'Faulty'. The Fully-Connected Encoder sacrifices the successful classification of a small number of 'Normal' samples in order to correctly identify each primary circuit break observed in this testing environment as 'Faulty'.

Classification of breaks located at the steam generation outlet was comparatively poor. This is perhaps due to the lower number of observed examples of secondary circuit breaks during training, or the potential for secondary circuit breaks to be harder to identify under the lower-fidelity of the secondary circuit physical model in comparison to the primary circuit. Only the Fully-Connected Encoder model is able to correctly classify any of these samples as 'Faulty', which suggests that secondary circuit breaks appear more similar to 'Normal' operational data from the perspective of these classifiers.

In the Target 1 domain, representing a slight shift in domain

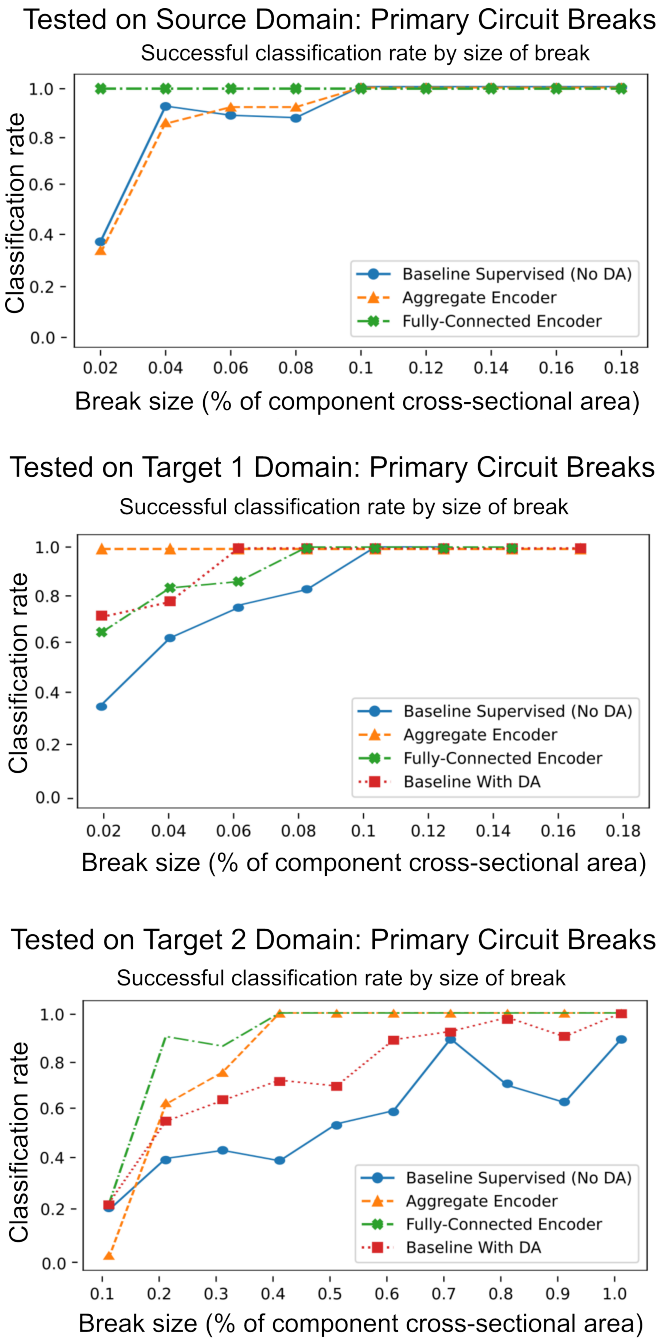

Figure 7. Successful classification rate by fault break size for each model combination, when tested on each domain.

distribution from the source domain, the Baseline Supervised model retained its ability to identify 'Normal' operational data, whilst 'Faulty' sample classification accuracy degraded. As in the Source domain, the Baseline Supervised and Aggregate Encoder models were not able to correctly identify any secondary circuit breaks as 'Faulty'. The Baseline With DA model (using $\mathcal{L}_{DA}$) gives some improvement in primary cir-

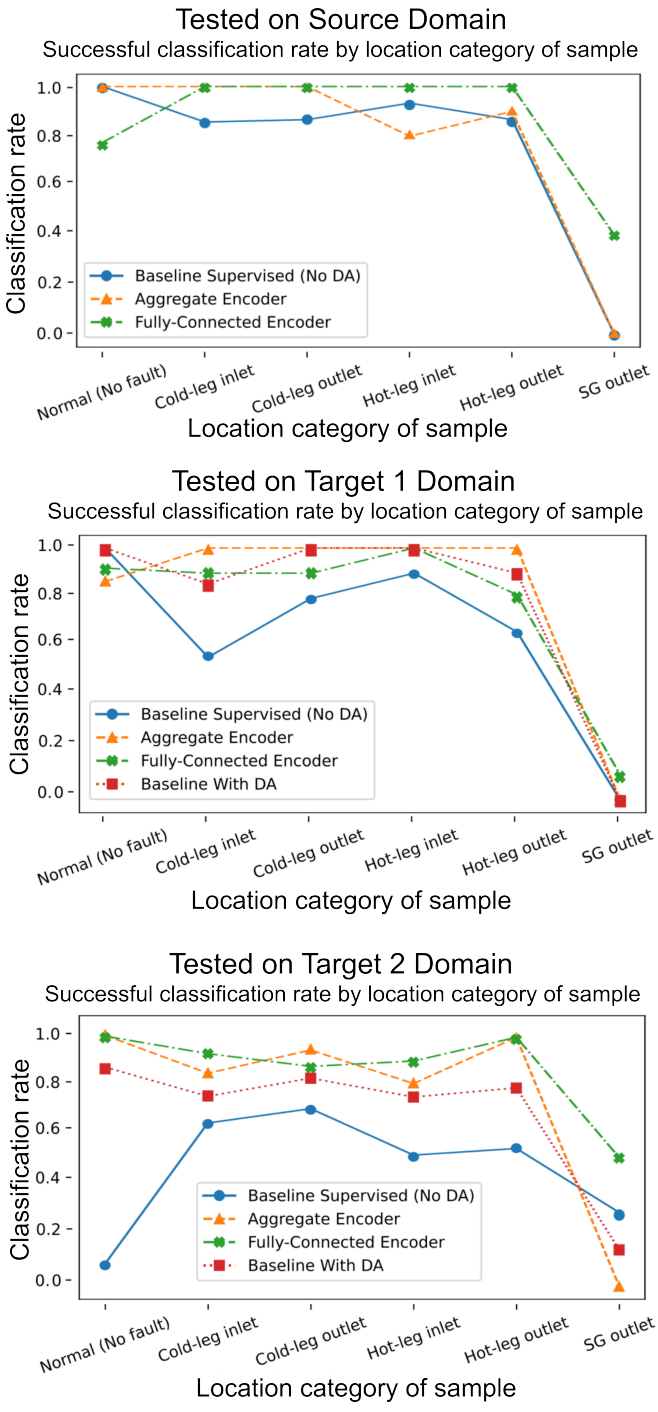

Figure 8. Successful classification rate by fault location for each model combination, when tested on each domain.

cuit break identification, but remains unable to recognise the 'Faulty' class of secondary circuit break samples. As in the previous testing domain, the Fully-Connected Encoder is the only model capable of correctly classifying any steam generator outlet samples, albeit a very small proportion of the samples it observed.

In the most extreme example of domain divergence between the Source domain and the Target 2 domain, the classification ability of the Baseline Supervised model, using $\mathcal{L}_S$, deteriorates. In this unfamiliar domain, the Supervised model is barely able to correctly classify any 'Normal' data. The Baseline model using $\mathcal{L}_{DA}$ restores some classification ability of 'Normal' samples, and improves the detection of 'Faulty' samples in all locations except the secondary circuit. Once again, the Fully-Connected Encoder displays the best detection performance for secondary circuit breaks. This indicates the importance of combining the sensor channels at the Encoder level in order to generate generalised features which can remain relevant between disparate training and testing domains.

5. Conclusion

The results detailed in this work highlight the value in incorporating domain adaptation techniques in scenarios where discrepancies exist between the training and testing data domains. Even when the scale of these discrepancies can become large, fundamental DA concepts provide a significant improvement in performance when compared to a conventional supervised learning approach. Additionally, the results draw attention to the importance of combining input sensor channels in this context. Models which consider information from multiple sensed sources simultaneously during their construction of each encoded feature map retained a much greater proportion of their classification ability seen in the Source domain classificaiton performance.

Acknowledgment

This work was partly funded by the Aerospace Technology Institute under the REINSTATE project.

References

Aldemir, T. (2013). A survey of dynamic methodologies for probabilistic safety assessment of nuclear power plants. *Annals of Nuclear Energy*, *52*, 113-124. (Nuclear Reactor Safety Simulation and Uncertainty Analysis)

Chen, J., Wang, J., Zhu, J., Lee, T. H., & de Silva, C. W. (2021). Unsupervised cross-domain fault diagnosis using feature representation alignment networks for rotating machinery. *IEEE/ASME Transactions on Mechatronics*, *26*(5), 2770-2781.

Farber, J. A., & Cole, D. G. (2020). Detecting loss-of-coolant accidents without accident-specific data. *Progress in Nuclear Energy*, *128*, 103469.

Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W.-K., & Yang, H. (2020). Status of research and development of learning-based approaches in nuclear science and engineering: A review. *Nuclear Engineering and Design*, *359*, 110479.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., &

Muller, P.-A. (2019, March). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963.

Lee, G., Lee, S. J., & Lee, C. (2021). A convolutional neural network model for abnormality diagnosis in a nuclear power plant. *Applied Soft Computing*, *99*, 106874.

Mahmoodi, R., Shahriari, M., Zolfaghari, A., & Minuchehr, A. (2011). An advanced method for determination of loss of coolant accident in nuclear power plants. *Nuclear Engineering and Design*, *241*(6), 2013-2019. ((W3MDM) University of Leeds International Symposium: What Where When? Multi-dimensional Advances for Industrial Process Monitoring)

Qian, Q., Qin, Y., Luo, J., Wang, Y., & Wu, F. (2023). Deep discriminative transfer learning network for cross-machine fault diagnosis. *Mechanical Systems and Signal Processing*, *186*, 109884.

Rivas, A., Delipei, G. K., Davis, I., Bhongale, S., & Hou, J. (2024). A system diagnostic and prognostic framework based on deep learning for advanced reactors. *Progress in Nuclear Energy*, *170*, 105114.

Sakurahara, T., O'Shea, N., Cheng, W.-C., Zhang, S., Reihani, S., Kee, E., & Mohaghegh, Z. (2019). Integrating renewal process modeling with probabilistic physics-of-failure: Application to loss of coolant accident (loca) frequency estimations in nuclear power plants. *Reliability Engineering & System Safety*, *190*, 106479.

Wang, H., jun Peng, M., Ayodeji, A., Xia, H., kun Wang, X., & kang Li, Z. (2021). Advanced fault diagnosis method for nuclear power plant based on convolutional gated recurrent network and enhanced particle swarm optimization. *Annals of Nuclear Energy*, *151*, 107934.

Wei, L., & Keogh, E. (2006). Semi-supervised time series classification. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (p. 748–753). New York, NY, USA: Association for Computing Machinery.

Zhang, J., Li, X., Tian, J., Jiang, Y., Luo, H., & Yin, S. (2023). A variational local weighted deep sub-domain adaptation network for remaining useful life prediction facing cross-domain condition. *Reliability Engineering & System Safety*, *231*, 108986.

Zhang, Y., Ren, Z., Feng, K., Yu, K., Beer, M., & Liu, Z. (2023). Universal source-free domain adaptation method for cross-domain fault diagnosis of machines. *Mechanical Systems and Signal Processing*, *191*, 110159.