Transfer learning methodology for machine learning based fault detection and diagnostics applied to building services

K Chavan, N Réhault, T Rist

Fraunhofer Institute for Solar Energy Systems, Freiburg-im-Breisgau, Germany
Email: nicolas.rehault@ise.fraunhofer.de, tim.rist@ise.fraunhofer.de

Abstract. Machine Learning (ML) models for Fault Detection and Diagnosis (FDD) can automatically detect anomalies in the operation in large facilities or district heating networks and can help tackling energy wastes. Nevertheless, the development of ML-models is a costly and tedious task requiring large amounts of labelled data. Setting up ML-models for a high number of systems is effort and know-how intensive. However, assets like commercial buildings and district heating networks are constituted of systems with similar topologies. Transferring a ML model initially trained on a source system to a multitude of similar target systems, can help reducing the training costs and facilitating the scalability of ML-based FDD in those assets. To enable this, we have developed a methodology that assesses the potential for Transfer Learning (TL) from a source system to target systems by determining the covariate and concept shifts between the source and target domains and integrating the source model into the target system if the TL assessment is positive. We used a patented method for the model development, that combines two ML-models, that are initially trained on a source system by means of a feedback system. We implemented this methodology on district heating (DH) substations, as DH systems typically contain this kind of subsystems with similar topologies and have thus a high scalability potential for TL. Initial findings showed the effectiveness of TL in adapting the source model to the target domain, resulting in enhanced FDD capabilities with significantly reduced training efforts.

1. Introduction

The operation of buildings and district heating or cooling networks is managed today by modern control systems designed for an optimal energy efficiency. Nevertheless, faults and anomalies can happen, resulting in both performance degradation and energy losses. Integrating Fault Detection and Diagnosis (FDD) methods in monitoring and maintenance processes of these systems can help identifying errors and flaws timely and efficiently and implementing energy conservation measures leading to significant energy savings [1]. FDD has been a large research field in the last decades and the first applications in the real estate and in the district heating industries have emerged together with the higher availability of measurement data and of web-based analytic services. FDD methods can be roughly divided into two categories: expert knowledge-based and Machine Learning (ML) methods. Expert knowledge-based methods include e.g. rule-based methods and physical-model based approaches. ML methods use measurement data and build an abstract representation of the analyzed systems. Contrary to expert-knowledge based methods, ML models require few prior knowledge of the system and are thus easy to setup if sufficient data is available [2]. Furthermore, they have the capacity to detect novelties. Commonly used ML methods for FDD include support vector machines, Bayesian networks, clustering analysis, decision tree or neural networks [3]. Nevertheless, ML methods require a large amount of labelled data of the normal and faulty system behavior to enable a reliable system supervision. As the

labelling and training process is time-consuming and require expert know-how, transferring a model, that has been pretrained on a source system to a large amount of similar target systems would enhance the scalability of ML for FDD in buildings. For example, in a large district heating (DH) system with hundreds of DH substations, transferring a model trained for FDD on a single DH substation defined as source system to a multitude of target DH substations, would significantly reduce the training efforts and enable reusing know-how gained on the source system.

Today, only few studies have investigated Transfer Learning (TL) for FDD for building energy applications. In [4], Liu et al. compared the performance of five models for FDD in building energy systems, including one trained from scratch on the source domain and four that made use of a pre-trained source model. In [3],Zhu et al. proposed an approach which uses a domain adversarial neural network trained on a source chiller to generate a diagnostic model for a target chiller. Martinez Viol et al. [5] emphasize the need for domain similarity analysis in TL for HVAC systems. They then present a data selection method and neural network models to tackle dissimilarity problems between datasets.

In our research, we opted for conventional ML models due to their ease of training compared to the deep neural networks mentioned in the previous studies. Despite their simpler architectures, these models delivered satisfactory results. [2]. We used here a patented method developed by Benndorf et al., which combines two ML models with a feedback system, used to support the model training [6]. The method, named COMETH, is based on two ML models used here for the classification of faulty and fault-free data, where the first model is chosen to have high sensitivity (here DBSCAN) and the second with a high specificity (here Decision Trees). In this paper, we present a methodology that enables assessing the possibility to transfer the source model trained with the above-mentioned method to a similar target system and then integrating the source ML-model in the target domain. In Chapter 2, we introduce the basic concepts related to TL. Chapter 3 provides a description of the developed methodology for TL evaluation and Chapter 4 details the integration approach of a pretrained model in a target domain. In chapter 5, we present the results on a use case with district heating substations. Chapter 6 concludes the paper with a short summary and discussion of our results.

2. Transfer learning definitions

A domain D is defined as the combination of an input space X, an output space Y and an associated probability distribution P. The input space X is constituted of time-series data from sensors of the analysed system also referred as features. The joint distributions between the input variables x and the output fault class y, are expressed as $P(x, y) = P(x|y)p(y)$ and $P(x, y) = P(y|x)p(x)$.

A domain shift is a change in the statistical distribution of data between the source and the target domains. It can be classified into three categories, prior shift, covariate shift and concept shift [7]:

1. *Prior shift*: The prior probabilities of the output classes are different $P_s(y) \neq P_t(y)$, but the conditional distributions are equivalent $P_s(y|x) = P_t(y|x)$. It means that the probability that a certain fault occurs is different between the source and target domains, but the causes for the occurrence of this fault are the same. A prior shift can for example occur in a DH substation after a maintenance task who might reduce the occurrence of faults [7].
2. *Covariate shift*: This kind of shift happens when the distribution of the features changes between source and target domains $P_s(x) \neq P_t(x)$, but the conditional probability distribution remains constant $P_s(y|x) = P_t(y|x)$ [7].
3. *Concept shift*: This shift occurs when the data distributions remain constant and the conditional distribution changes between source and target domains change $P_s(y|x) \neq P_t(y|x)$. For instance, in the context of FDD, the underlying reasons for the occurrence of faults may vary between the source and target domains [7].

3. Methodology

The main objective of the developed methodology described below is to assess the feasibility of TL by evaluating the magnitude of the domain shift. Pre-trained models, which have been fine-tuned with user feedback in one single source system are then transferred for FDD on a multitude of target systems, while minimizing the learning efforts through feedback for the target systems. By assessing the domain shift, we can focus the efforts on implementing TL only for systems, for which it is feasible. It enables thus saving time and having an efficient process. We specifically focus on examining systems for covariate shift and concept shift and did not consider prior shift in our work because it involves comparing distributions of fault labels in the source and target domains, and we lacked per se labelled data of faulty behaviour in the target domains.

In the following, we describe the different steps of our methodology, which are illustrated in Figure 1.

Step 1: In this initial step, common features for all the analyzed systems are selected and the features units and semantic descriptions are standardized to ensure consistency. Physically implausible outliers are removed from the time-series data of the features.

Step 2: Pairs of all possible source-targets combinations are established.

Step 3: The source system is selected as the system exhibiting the highest similarity coefficient, as defined in equation 1, with the greatest number of target systems.

$$S_c = \frac{\mathrm{d(R1)-d(R1,R2)}}{\mathrm{d(R1)}} \times 100\ [\%] \qquad (1)$$

Where, d(R1) and d(R1, R2) are the Euclidean distances between the correlation matrices of the features of the source system and of the source and the target system [8].

Step 4: The similarity coefficient Sc threshold is set. Initially, an empirical threshold value of 80% is chosen.

Step 5: The concept shift is then evaluated based on the Sc value If Sc is higher than the Sc threshold, then TL is considered feasible.

Step 6: Among pairs of source and target system for which TL is initially feasible, the pair with the lowest S_c is used to determine an initial maximum mean discrepancy (MMD) threshold [9]. The MMD is a statistical test to determine the differences between the probability distributions of the source and target domains, which we use here to evaluate the covariate shift and we calculate it here for all corresponding features of the source-target pairs.

Step 7: The source model is trained in the source domain by using the combined method COMETH with feedback system [3].

Step 8: The source model, consisting of the pipeline including the standardization routines, the actual ML-models algorithm with their hyper parameters and the training data of the source domain, is transferred to the target domain. The training data of the source model underwent Z-score standardization, and it was subsequently adapted with each new application data from the target domain, which was obtained through user feedback.

Step 9: The classification performance of the target domains models with and without TL is calculated by means of following classification metrics: sensitivity, precision, and number of feedback samples. A feedback sample being feedback from a single timestamp.

Step 10: If the performance is positively evaluated by the users, the target models can be used perform FDD on the target systems. The MMD and Sc thresholds can then be reused for additional target systems.

Step 11: If the performance is not satisfactory, features with high MMD value are removed and the MMD threshold is updated to the new highest value of MMD among all features. The process is repeated

from Step 6. It is to be noted that the removal of some features can lead to important information losses, which hamper the setup of good models.

Step 12: The Sc thresholds can then be reduced stepwise until the user can transfer the source model to many target systems while keeping a satisfactory FDD performance.

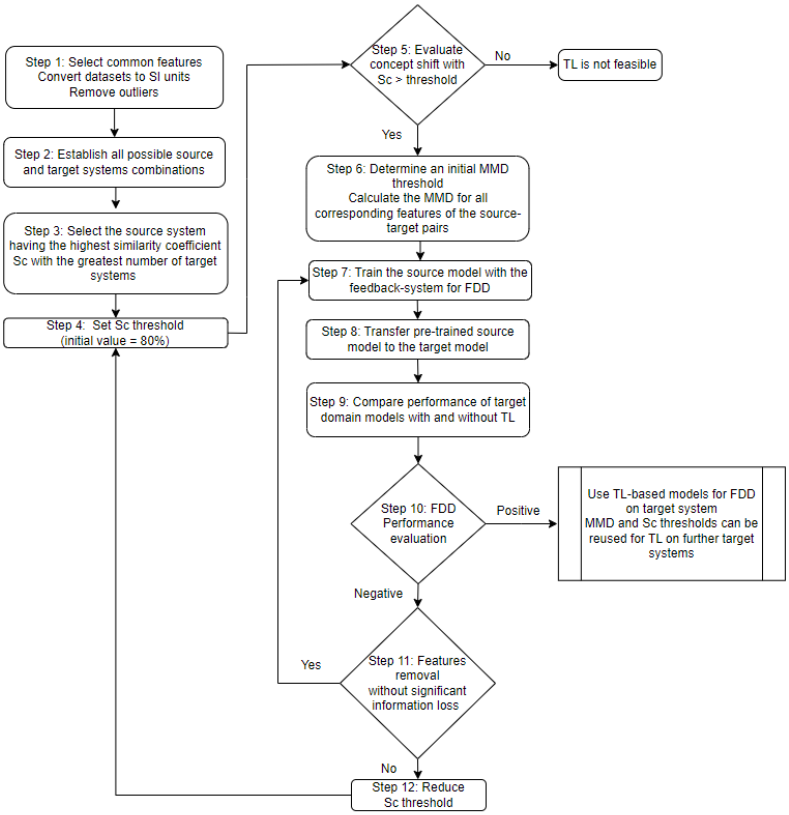

Figure 1. Transfer Learning process diagram [1]

This methodology allows users to establish their own performance standards for TL by determining stepwise the Sc and MMD thresholds expressing respectively the concept and covariate shifts.

4. Use cases

We considered a small district heating network located in the South of Germany to demonstrate our methodology. Data between January and March 2021, with a 5 min sampling rate, was available for 4 substations with heating powers ranging from 50 to 80 kW. Each substation provided the following features: secondary side supply temperature, secondary side space heating supply and return temperatures, primary side supply and return temperatures, domestic hot water (DHW) heating power, space heating power. Our methodology enabled selecting one substation, training it as source system, assessing the TL feasibility for the three remaining substations and transferring the source model in the target domains. The TL feasibility with the remaining substations was reached iteratively by removing the DHW heating power as it exhibited a high variation between the substations. This result shows that features related to specific operation pattern of one system, like in our use case the DHW consumption, can lead to a decrease in the performance of TL due to high distributional differences. Our methodology enables identifying such features and excluding them if the information loss is acceptable. To test our algorithms, synthetic faults causing drops in the hot water supply temperature were introduced in the source and target domains. The source model and the target models of the three target substations without TL were initially trained on two weeks of data from January 1, 2021, to January 15, 2021, and

the application phase last until March 4. 2021, corresponding thus to a representative operation in a heating season. The application phase with TL started directly on January 1, 2021. The detailed performance indicators are provided in Table 1. Figure 2 depicts the mean values of the performance indicators of the three target systems with and without TL

Table 1. FDD performance of the three target systems without and with TL

Source system	Target system	without Transfer learning								with Transfer learning						
		Model	TP	FP	TN	FN	Sensitivity	Precision	Feedback samples	TP	FP	TN	FN	Sensitivity	Precision	Feedback samples
DH substation 7	DH substation 5	DT	54	0	14245	101	0,35	1,00	36	119	24	18288	1	0,99	0,83	25
		DBSCAN	51	75	14202	72	0,41	0,40	114	84	0	18312	36	0,70	1,00	36
DH substation 7	DH substation 6	DT	84	1	14241	74	0,53	0,99	51	108	10	18314	0	1,00	0,92	10
		DBSCAN	132	18	14248	2	0,99	0,88	20	94	18	18306	14	0,87	0,84	32
DH substation 7	DH substation 8	DT	62	2	14251	85	0,42	0,97	32	117	38	18273	4	0,97	0,75	42
		DBSCAN	74	29	14243	54	0,58	0,72	47	110	0	18311	11	0,91	1,00	11

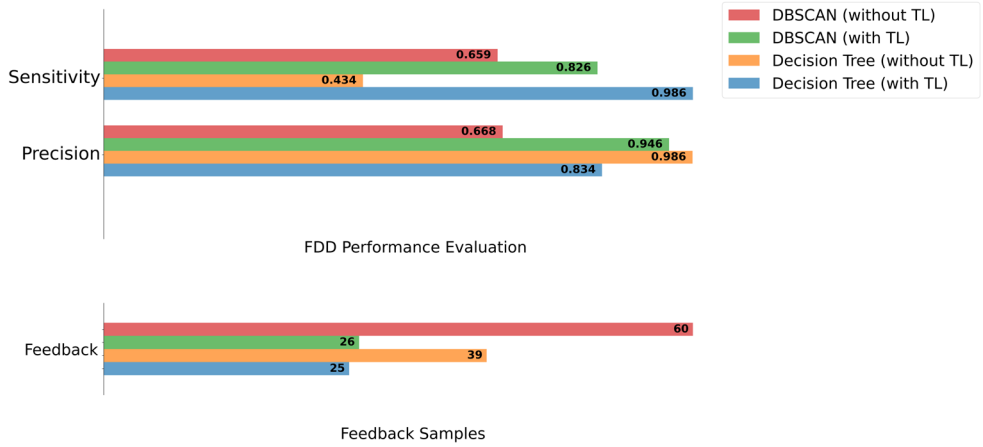

Figure 2. Average FDD performance of the target models with and without transfer learning

TL significantly improved the average sensitivity of both DBSCAN and DT models by 25.4% and 127.5% respectively, compared to models without TL. The precision of DBSCAN models increased by 41.6%, while DT models experienced a decrease of 15.4%. The increase in sensitivity with TL can be attributed to the acquired knowledge of fault characteristics from the source domain through iterative training with the COMETH method, which is then transferred to the target domain via TL. This is particularly interesting for DT models, as their sensitivity depends on the number of fault types learned during the training phase with labeled data. TL enables DT models to access a larger labeled dataset, leading to a significant increase in sensitivity. The drop in precision observed in DT models with TL is a result of the higher number of fault predictions made by the models without TL, which resulted in high precision but low sensitivity, as many faults went undetected. After applying TL, the number of feedback samples could be reduced by 57% for DBSCAN and 36% for DT. These preliminary findings demonstrate the effectiveness of TL in adapting the source model to the target domain, resulting in enhanced fault detection capabilities with significantly reduced training efforts.

5. Conclusion and outlook

We have introduced a methodology developed to facilitate the decision-making process involved in transferring a pre-trained ML model from one system to another. The objective is to improve the scalability of ML models for FDD in systems like buildings or heating networks constituted of a multitude of similar subsystems. For the learning and fine-tuning of the models, we utilized a patented combination of ML techniques, along with a feedback system. We combined here decision trees with the clustering algorithm DBSCAN. To demonstrate the effectiveness of our methodology, we applied it to substations within a small district heating system. By calculating the similarity coefficient between

the distributions of all possible source-target systems combinations, we were able to efficiently identify the most suitable source system. We could exemplarily show on a small-scale system, that our approach allows for the transfer of knowledge from the source domain to the target domains, eliminating the need for a training phase in the target domains in implementing FDD. Additionally, it minimizes the efforts required for fine-tuning in the target domains. In the case of systems with same sizes but different operating principles, the covariate shift is likely to be minimal, but the behavior differences of the systems can result in a high concept shift, which can lead to poor performance of the pre-trained model in the target domain. On the other hand, in cases where the systems have different sizes but similar operating principles, the concept shift is likely to be low due to similar operating behaviors, but the features related to the size of the system like e.g., mass flows and thermal or electrical power lead to a high covariate shift. Through the evaluation of the concept and covariate shifts, our methodology supports the user, at removing features that hampers the transferability of the model pre-trained in the source domain. Nevertheless, only the user can assess the information loss that a feature removal induces. In future works, we will investigate different types of systems, such as chillers, air handling units or heat pumps with the objective to consolidate our results and to determine standard feature sets for typical systems in buildings and large heating or cooling networks. Furthermore, we plan using optimization techniques to mathematically determine the threshold set today empirically for the concept shift evaluation.

Acknowlegements
This work was funded by the German Federal Ministry of Education and Research in the project KETEC under grant number 03SF0623C.

References

[1] F. Zhang, N. Saeed, and P. Sadeghian, “Deep learning in fault detection and diagnosis of building HVAC systems: A systematic review with meta analysis,” *Energy and AI*, vol. 12, p. 100235, 2023, doi: 10.1016/j.egyai.2023.100235.
[2] G. A. Benndorf, D. Wystrcil, and N. Réhault, “Energy performance optimization in buildings: A review on semantic interoperability, fault detection, and predictive control,” *Applied Physics Reviews*, vol. 5, no. 4, p. 41501, 2018, doi: 10.1063/1.5053110.
[3] X. Zhu, K. Chen, B. Anduv, X. Jin, and Z. Du, “Transfer learning based methodology for migration and application of fault detection and diagnosis between building chillers for improving energy efficiency,” *Building and Environment*, vol. 200, p. 107957, 2021, doi: 10.1016/j.buildenv.2021.107957.
[4] J. Liu *et al.,* “Transfer learning-based strategies for fault diagnosis in building energy systems,” *Energy and Buildings*, vol. 250, p. 111256, 2021, doi: 10.1016/j.enbuild.2021.111256.
[5] Victor Martinez-Viol, Eva M. Urbano, Jose E. Torres Rangel, and Miguel Delgado-Prieto and Luis Romeral, “Semi-Supervised Transfer Learning Methodology for Fault Detection and Diagnosis in Air-Handling Units,”
[6] G. A. Benndorf, D. Wystrcil, and N. Réhault, “A fault detection system based on two complementary methods and continuous updates,” *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 353–358, 2018, doi: 10.1016/j.ifacol.2018.09.601.
[7] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A Brief Review of Domain Adaptation,” Oct. 2020. [Online]. Available: http://arxiv.org/pdf/2010.03978v1
[8] Mihail N. Kolountzakis, Kiriakos N. Kutulakos, “Fast computation of the Euclidian distance maps for binary images,” vol. 43, 181-184, 1992.
[9] L. Ouyang and A. Key, “Maximum Mean Discrepancy for Generalization in the Presence of Distribution and Missingness Shift,” Nov. 2021. [Online]. Available: http://arxiv.org/pdf/2111.10344v2