Utilizing Autoencoders to Improve Transfer Learning when Sensor Data is Sparse

Maira Alvi
The University of Western Australia
Perth, Australia
maira.alvi@research.uwa.edu.au

Rachel Cardell-Oliver
The University of Western Australia
Perth, Australia
rachel.cardell-oliver@uwa.edu.au

Tim French
The University of Western Australia
Perth, Australia
tim.french@uwa.edu.au

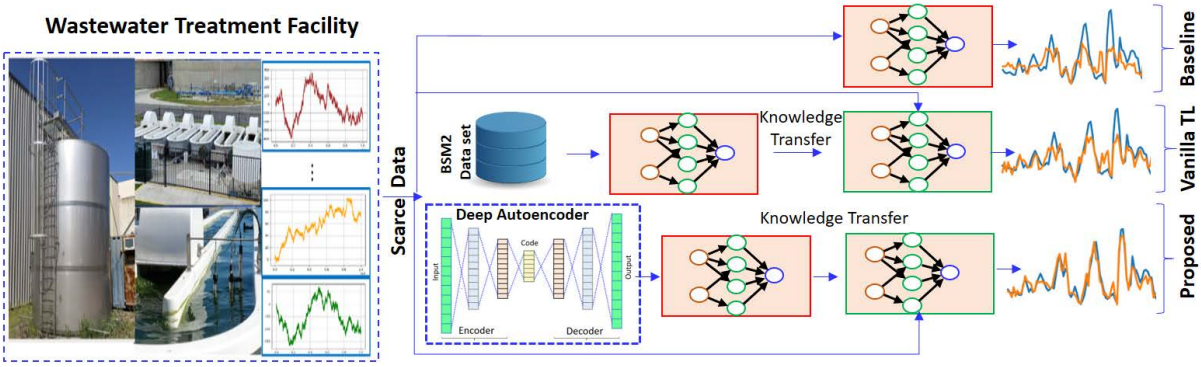

Figure 1: Comparison between a traditional deep learning baseline (top), transfer learning (middle) and our approach for enhanced transfer learning (bottom). The line plots (right) indicate that using an autoencoder to generate training data leads to more accurate sensor predictions (orange) for ground truth (blue).

ABSTRACT

Accurate deep predictive models of wastewater processing plants are important to ensure operational parameters are safe and sustainable. Training such predictive models requires large volumes of data that is hard to find in the wastewater domain. Transfer learning addresses the problem, by training predictive models using data from an adopted domain, and fine-tuning it on the target domain. However, due to the significant distributional shift between the commonly adopted source domains for transfer learning and the target domain of wastewater processes, transfer learning rarely performs at an acceptable level. This paper proposes a method to generate large volumes of training data with a similar distribution to a sample taken from the target domain, to boost transfer learning performance for the task, referred to as *AETL*. It leverages an autoencoder that systematically augments the target domain data such that synthetic samples it generates closely follow the target domain distribution. The results on the real-world data set establish the efficacy of the proposed framework.

CCS CONCEPTS

• **Applied computing** → *Physical sciences and engineering*.

BuildSys '22, November 9–10, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9890-9/22/11…$15.00
https://doi.org/10.1145/3563357.3567407

KEYWORDS

Wastewater Treatment, Autoencoder, Data Augmentation, Transfer Learning, Deep Learning

ACM Reference Format:
Maira Alvi, Rachel Cardell-Oliver, and Tim French. 2022. Utilizing Autoencoders to Improve Transfer Learning when Sensor Data is Sparse . In *The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3563357.3567407

1 INTRODUCTION

Urbanization and industrialization generates wastewater that can pose environmental and ecological challenges [9]. A wastewater treatment facility safeguards human health and preserves the ecosystem. These facilities monitor key parameters such as Ammonia and Nitrite in the discharge water. However, these key indicators of the process's state can be hard-to-measure because they require high-cost sensors or offline laboratory analysis [2]. As a result, only small data sets of measurements may be available for decision-making.

Soft sensing is a technique for estimating hard-to-measure parameters. It leverages easy-to-measure sensor readings to train a Deep Learning (DL) model to estimate the values of hard-to-measure indicators. However, a challenge for soft sensing in the wastewater domain is the lack of data for training DL models: often, there are insufficient relevant sensor readings due to the high associated cost of sensors and long associated delay of lab analysis.

Transfer Learning (TL) can be used with DL models when there is limited training data. In TL we transfer the knowledge gained from other problems (source domains) to assist the problem at hand

(target domain) where the available data is scarce. TL is useful in scenarios with a negligible covariant shift between source and target domains. This generally does not hold for wastewater processes due to significant difference in underlying processes and thus renders TL ineffective and handicap the performance of the overall DL models.

This paper proposes a method for synthesizing high-quality training data using autoencoders. We use the synthesized data for training a source model. Finally, the model is fine-tuned with the measured target domain data to train a generalizable DL model. We evaluate our technique on challenging real-world scarce data sets of a wastewater treatment process known as High Rate Algae Ponds (HRAPs). We show that this approach gives performance gain of 18% and 26% in terms of RMSE and 50%, and 20% on R^2 for nitrite and ammonia over existing methods.

2 RELATED WORK

Urban planning and smart cities literature widely use the term cold-start, which implies 'how to develop smart cities with limited data?' [11]. To mitigate the data scarcity problem, smart city practitioners can employ transfer learning (TL). In [4, 8] conventional TL is employed to train a source model using other available data sets and afterwards fine-tune over the target domain. Pisa et al. explored TL with the aim of minimizing training time associated with the design of the control loop in WWTPs [10]. Their work does not particularly focus on data scarcity issues and exploits synthetic samples generated from a mechanistic model Benchmark Simulation Model (*BSM*).

In comparison, this paper takes a fundamentally different approach. We propose a technique that remedies the data scarcity problem by generating training data to minimize the distance between the source and target domain data. The data we generate assists transfer learning.

3 PROBLEM FORMULATION

Let $\mathcal{X}$ denote the distribution of sensor time-series data captured from a Wastewater Treatment Plant (WWTP). The data set, $\boldsymbol{X}$, includes 'm' samples of 'n' different process variables i.e., $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}^m \in \mathbb{R}^{m \times n}$ s.t., $\boldsymbol{x}_i \sim \mathcal{X}$. Here, each sample from $\boldsymbol{X}$, i.e., $\boldsymbol{X}_i \in \mathbb{R}^n$ has an associated variable $\boldsymbol{y}_i$, where $\boldsymbol{y}_i \sim \mathcal{Y}$. Although both $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$ indicate the underlying variables of the WWTP, we have deliberately used distinct variables for ease of formulation. Here, $\boldsymbol{y}_i$ are the hard-to-measure and $\boldsymbol{X}_i$ easy-to-measure process parameters. Our goal is to train a deep learning model $\mathcal{M}(\boldsymbol{X}_i, \boldsymbol{y}_i; \boldsymbol{\Theta})$, with learn-able parameters $\boldsymbol{\Theta}$ that can predict $\boldsymbol{y}_i$ with the aid of only $\boldsymbol{X}_i$'s. The training objective can be casted as optimization of $\boldsymbol{\Theta}$ using some loss function '$\mathcal{L}(.)$' to encode the distribution $\mathcal{X}$, and $\mathcal{Y}$. Although it is theoretically possible to achieve it with the aid of a very large number of samples from $\mathcal{X}$, and $\mathcal{Y}$. However, in reality, the scarce availability of training data, $\mathcal{M}(.)$ struggles to model $\mathcal{X}$, and $\mathcal{Y}$ faithfully.

Transfer learning has demonstrated performance improvements arising from data scarcity issues in several domain problems [12]. It aims to utilize the knowledge gained over some source domain (s), where a large number of training samples are available, to learn the mapping to a target domain where samples are scarce. Thus, $\boldsymbol{\Psi} : \hat{\mathcal{M}}(\boldsymbol{X}_i^s, \boldsymbol{y}_i^s; \hat{\boldsymbol{\Theta}}) \rightarrow \mathcal{M}(\boldsymbol{X}_i, \boldsymbol{y}_i; \boldsymbol{\Theta})$, where $\boldsymbol{X}_i^s \sim \mathcal{X}^s, \boldsymbol{y}_i^s \sim \mathcal{Y}^s$ are large number of samples from a source domain and $\boldsymbol{X}_i, \boldsymbol{y}_i \subset (\boldsymbol{X}, \boldsymbol{Y})$ are small subset of available samples in a target domain $\mathcal{X}$ and $\mathcal{Y}$, while $\hat{\mathcal{M}}(.)$ denotes a pre-trained model (source domain), '$\rightarrow$' indicates transfer learning and $\mathcal{M}(.)$ is the final fine-tuned model over the target domain. The performance of model $\mathcal{M}(.)$ is highly contingent over mapping $\boldsymbol{\Psi}$ and is mainly governed by the distributional shift between the source and target domains i.e., $||\mathcal{X}^s - \mathcal{X}||$ and $||\mathcal{Y}^s - \mathcal{Y}||$. A smaller distributional shift is desired to enable effective transfer learning so that $\mathcal{M}(.)$ becomes more representative of target domain distribution.

Bottleneck : Wastewater domain suffers from two problems that have hindered the adoption of transfer learning approaches. Firstly there are only a handful of data sets that are available in the public domain [2], and secondly, these data sets have a significant distributional shift. This is due to the fact that each of the data sets is based on an entirely different treatment process that is governed by different biological and physio-chemical principles that lead to significantly different distributions. For instance, activated sludge and anaerobic digestion process *BSM2* [6], and HRAP [2]. Consequently, a source domain data can be almost irrelevant for a target domain and thus can compromises in possible performance gain from the mapping $\boldsymbol{\Psi}$ to mitigate the data scarcity issues in a target domain.

Motivation : With prudent consideration of the above-mentioned issues in terms of (1)distributional shift between source and target, (2) making $\boldsymbol{X}, \boldsymbol{Y}$ more representative of the distribution $\mathcal{X}, \mathcal{Y}$. We propose a method inspired by deep identity functions (autoencoders) to generate synthetic data closely following a target domain distribution. Namely, we let $\boldsymbol{\Psi} : \overline{\mathcal{M}}(\boldsymbol{X}_i^{ae}, \boldsymbol{y}_i^{ae}; \overline{\boldsymbol{\Theta}}) \rightarrow \mathcal{M}(\boldsymbol{X}_i, \boldsymbol{y}_i; \boldsymbol{\Theta})$. Here, $\boldsymbol{X}_i^{ae} \sim \mathcal{X}^{ae}, \boldsymbol{y}_i^{ae} \sim \mathcal{Y}^{ae}$ is synthetic data samples generated from autoencoder, such that $||\overline{\boldsymbol{\Theta}} - \boldsymbol{\Theta}|| << ||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||$. Thus, we aim to reduce the distributional shift between a source and target domain in the context of wastewater treatment processes and then transfer of knowledge from $\overline{\mathcal{M}}(.)$ to $\mathcal{M}(.)$ to promote higher generalization. Our results in Section 5 verify this.

4 PROPOSED APPROACH

Distinct from the aforementioned transfer learning techniques, we propose a method to generate synthetic data systematically. It utilizes deep autoencoders that are trained with the aim to closely mimic the distribution of target domain data. The minimal covariate shift between source and target domain enables the effective transfer of knowledge and thus yields more robust and accurate predictive models. Detailed empirical evaluation of the proposed technique demonstrates that it leads to significant performance gain over vanilla transfer learning under a similar setting. We describe the proposed technique, named Autoencoder Transfer Learning (*AETL*) below.

Our method utilizes a deep autoencoder model and is composed of two modules, namely an encoder and a decoder. The encoder includes vanilla LSTM layer that projects input features ($\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m \times n}$) to a latent code $\boldsymbol{Z} \in \mathbb{R}^{m \times o}$, s.t., $o << n$. Following this, the decoder part of network endevours to reconstruct the input data points ($\boldsymbol{X}^{ae}, \boldsymbol{Y}^{ae}$) $\in \mathbb{R}^{m \times n}$ from information available in latent space. Thus, the goal of autoencoder is to minimize distributional difference between original and reconstructed data points. Following the standard convention, in all of experiments, size of latent

space was kept as half of the input features size, the loss function was kept as 'mean squared error' loss and 'Adam' optimizer was utilized with maximum number of training epochs as 1000 with early stopping. Once the model has learned the identity mapping, synthetic data was generated by sampling, using the decoder to generate additional time steps data. The presented results include the results where the training data was augmented with synthetic data (3× the real data).

The similarity in probability distribution of source $(\mathcal{X}^{ae}, \mathcal{Y}^{ae})$ and target $(\mathcal{X}, \mathcal{Y})$, data can be illustrated with the aid of Kullback-Leibler (KL) divergence [7]. It is a standard metric to measure the divergence between two probability distributions (relative entropy) and aids in evaluating the similarity in distributions. Figure 2 illustrates difference between two data generated with proposed method and the commonly adopted source domain (e.g., *BSM2*). Here, a smaller value of KL divergence is analogous to a closely matching distribution and thus facilitates effective transfer learning. Following generation of data for the source, a DL model is trained over it ($\overline{\mathcal{M}}(.)$), afterwards this model is fine-tuned over available limited target domain data ($\mathcal{M}(.)$). This will be referred to as 'Autoencoder transfer learning' (*AETL*) in the text.

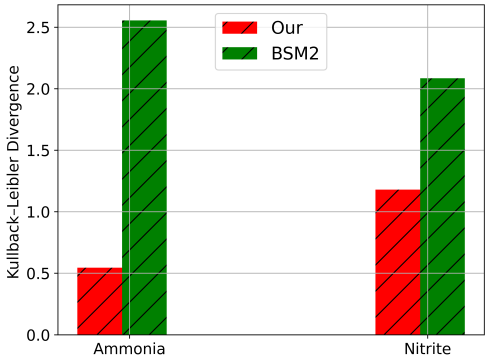

Figure 2: Comparison of KL Divergence (lower is better) for Nitrite and Ammonia between the source domain data and target domain for our autoencoder approach and *BSM2* mechanistic model data.

5 RESULTS

We first provide details of two data sets and our experimental setup, followed by the results.

HRAP Data set: The empirical study leverages real-world data set from an anaerobic-algae wastewater treatment process. It is a two-stage treatment process composed of an Up-flow Anaerobic Sludge Blanket (UASB) reactor that decomposes organic pollutants while producing bio-gas as a by-product. Following digestion, water is then ingested by algae raceways, also known as High Rate Algae Ponds (HRAPs), that remove nutrients and produce micro-algae as a by-product. Before the final discharge of treated water (effluent) into the nature, various parameters are actively monitored in HRAPs using real-time sensor systems as well as laboratory analysis. This study has utilized HRAP data set acquired between 10 August 2020 and 30 July 2021 via both sensor and laboratory analysis.

The sensor systems provide online measurements of pH, temperature, and dissolved oxygen DO at 15-min intervals. Whereas the offline measurements of Ammonia and Nitrite (key indicators) are performed daily in laboratories. Due to the associated cost of lab analysis, these parameters are an ideal test bed for development of soft sensors. Ammonia and Nitrite are key indicators that are strictly regulated in discharged water streams. Based on their significance, this study surrogates them using other online measurements. For the preparation of data to train DL model, the available data from sensors was filtered such that it resulted in 1 to 1 correspondence between online and offline samples. It is worth mentioning data over this long duration of almost a year only has ≈ 204 samples and is referred as target domain data set. It is important to note that the very limited nature of real samples makes training of DL models a challenging problem.

BSM2 Data set [6]: The benchmark simulation is an environment that defines plant process, influent loads, performance criteria and test procedures. It is the most recent version of BSM series and a plant-wide model of activated sludge (ASM1 [5]) and anaerobic digestion (ADM1[3]) process, available in the public domain. To generate effluent data the simulation was executed in two stages. Firstly, the simulation was performed in open-loop mode for 500 days using the constant influent data set to converge the system to a realistic starting state. Secondly, the simulation is performed in closed-loop mode for 609 days using the dynamic influent to generate the effluent data.

We utilize the final closed-loop for 609 days data with sampling frequency of 15 minutes. Under the above settings, 58464 samples of data were generated. The effluent parameters S_{NH} gN/m^3 (Ammonia and Ammonium), S_{NO} gN/m^3 (Nitrate and Nitrite), temperature, dissolved oxygen $S_o(g - CODm^{-3})$ and alkalinity are used to train a source model. Note that, this data set is used as a source domain data for vanilla transfer learning approach.

Experimental Setup: The effectiveness of our technique (*AETL*) is demonstrated by comparing it against two different conventional approaches namely; *Baseline* and $BSM2 - TL$. In *Baseline* the DL model is trained with the limited available target domain samples, while $BSM2 - TL$ first trains a source model over *BSM2* data and afterwards it is fine-tuned over the target domain data. Note that, '$BSM2 - TL$' is also known as vanilla transfer learning approach.

Implementation Details : In all the methods, the DL models were trained in a similar fashion. The source model is composed of 5 fully connected layers with 64, 64, 16, 16, 1 neurons in each layer and following standard convention the non-linearity was selected as *ReLU* function. These experiments were performed in the TensorFlow [1] framework in Python language. For all the models, the maximum training epochs were set as 200, batch size as 32 with 10-fold cross-validation. Learning rate was set as 0.0010 and ADAM optimizer was utilized for learning model parameters with a MSE loss function. For fine-tuning, two layers (near to input) were freezed and learning rate was reduced to 0.00010. These hyper-parameters were manually selected to train best performing models.

Evaluation Metrics : We report the results on the three standard evaluation metrics namely; Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Correlation of Determination (R^2).

Results : The comparative results of the conventional and proposed approache is summarized in Table 1. It can be observed that for ammonia prediction, *AETL* provides a performance gain of 26.52%, 47.54%, and 25% in terms of RMSE, MAE and R^2 over the closet

competitor. Similarly, for nitrite, *AETL* has superior performance with gains of 18.12%, 34.14%, and 50% respectively. Here, it is important to note that $BSM2-TL$ has sub-optimal performance than the original baseline. We conjecture that it may be the result of distributional shift between the source and target domain. Even though TL provide an effective prospect in case of limited data, however, it is only useful when there is somewhat smaller covariate shift. It is also apparent from the results that *AETL* outperforms its competitors because its main strength is derived from the synthetic data generated by the auto-encoder that has a small divergence between the source and target domain.

Table 1: Prediction results of Ammonia and Nitrite reported as (mean±std) on 10-fold cross-validation. The first column indicates the models used for comparison, evaluated on three metrics RMSE, MAE, and R^2. Best performing model results are highlighted in bold.

Model	Ammonia $NH_4-N(mg/L)$		
	RMSE	MAE	R^2
Baseline	2.619±0.653	1.806±0.348	0.609±0.229
$BSM2-TL$	3.021±0.727	2.064±0.374	0.461±0.246
AETL	**2.070±0.379**	**1.225±0.216**	**0.801±0.091**
	Nitrite $NO_2-N(mg/L)$		
Baseline	0.88365±0.06378	0.555±0.046	0.203±0.327
$BSM2-TL$	1.008±0.144	0.595±0.057	0.303±0.366
AETL	**0.745±0.149**	**0.413±0.052**	**0.660±0.234**

The additional visualization of R^2 score are included in Figure 3 that indicates the goodness of model's approximation of original values. It clearly indicate that proposed approach outperforms other methods and hence justifies significant performance gains. Further, the qualitative results of our approach are availalble in Figure 4. It illustrates that the predictions of *AETL* are precise and closely follow the original time series, with only slight occasional temporal misalignment. Overall, empirical and qualitative results ascertain the effectiveness of the proposed approach.

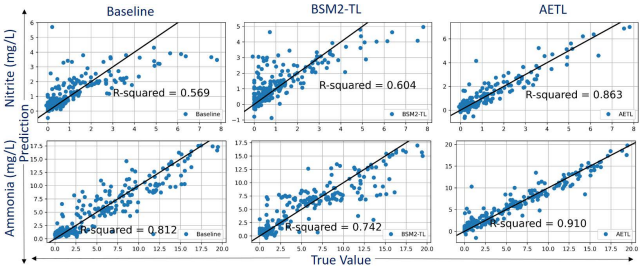

Figure 3: Scatter plots of prediction and ground truth for Nitrite and Ammonia for the *Baseline*, $BSM2-TL$ and *AETL*.

6 CONCLUSION

This work highlights that performance of TL relies on the distributional shift between the source and target domain. It proposes a novel yet simple method for effective TL for predictive modelling of wastewater key parameters. It trains an LSTM autoencoder with the aim of reducing the distributional difference between the source and target domain. This leads to significant performance improvements as validated by experiments using challenging real-world data sets of wastewater treatment processes. The proposed approach can be generalised for further applications in urban sensing.

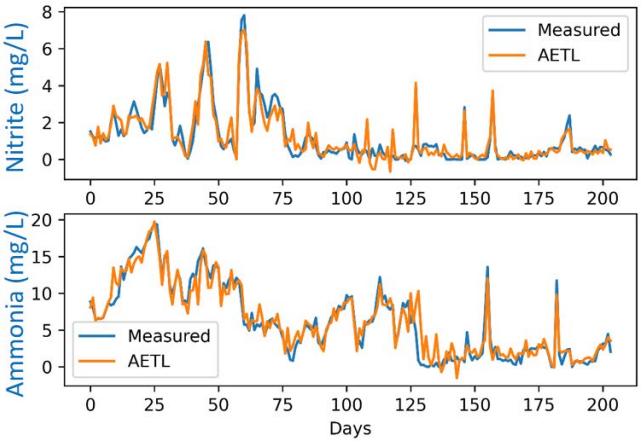

Figure 4: Qualitative prediction results of *AETL* for Nitrite and Ammonia. Ground truth time series labelled as 'Measured' is shown for reference.

Acknowledgments This research was supported by the Cooperative Research Centre Projects (CRC-P) 'Transforming wastewater treatment in regional Australia with robust technology for multiple benefits' under Grant CRCPSIX000079. The authors would like to thank Dr. Andrew Ward for the HRAP data set.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Maira Alvi, Tim French, Rachel Cardell-Oliver, Philip Keymer, and Andrew Ward. 2022. Cost Effective Soft Sensing for Wastewater Treatment Facilities. *IEEE Access* 10 (2022), 55694–55708. https://doi.org/10.1109/ACCESS.2022.3177201

[3] Damien J Batstone, J Keller, Irini Angelidaki, SV Kalyuzhnyi, SG Pavlostathis, A Rozzi, WTM Sanders, HA Siegrist, and VA Vavilin. 2002. The IWA anaerobic digestion model no 1 (ADM1). *Water Science and technology* 45, 10 (2002), 65–73.

[4] Hongchao Cheng, Yiqi Liu, Daoping Huang, Chong Xu, and Jing Wu. 2020. A novel ensemble adaptive sparse Bayesian transfer learning machine for nonlinear large-scale process monitoring. *Sensors* 20, 21 (2020), 6139.

[5] Mogens Henze, CP Lesley Grady, Willi Gujer, GvR Marais, T Matsuo, et al. 1987. Activated sludge model n 1. *IAWQ Scientific and Technical Report* 1 (1987).

[6] Ulf Jeppsson, M-N Pons, Ingmar Nopens, J Alex, JB Copp, KV Gernaey, Christian Rosén, J-P Steyer, and PA Vanrolleghem. 2007. Benchmark simulation model no 2: general protocol and exploratory case studies. *Water Science and Technology* 56, 8 (2007), 67–78.

[7] Solomon Kullback. 1987. The kullback-leibler distance. (1987).

[8] Pedro Oliveira, Bruno Fernandes, Cesar Analide, and Paulo Novais. 2021. Forecasting energy consumption of wastewater treatment plants with a transfer learning approach for sustainable cities. *Electronics* 10, 10 (2021), 1149.

[9] Chang Peng, Wang Kai, Zheng Kun, and Meng Fanchao. 2022. Monitoring of wastewater treatment process based on multi-stage variational autoencoder. *Expert Systems with Applications* 207 (2022), 117919.

[10] Ivan Pisa, Antoni Morell, Jose L Vicario, and Ramon Vilanova. 2021. Transfer Learning Approach for the Design of Basic Control Loops in Wastewater Treatment Plants. In *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 1–8.

[11] Leye Wang, Bin Guo, and Qiang Yang. 2018. Smart city development with urban transfer learning. *Computer* 51, 12 (2018), 32–41.

[12] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.