Supplementary Information

A Novel Deep Learning Framework with Variational Auto-Encoder for Indoor Air Quality Prediction

Qiyue Wu[a], Yun Geng[a], Xinyuan Wang[a], Dongsheng Wang[b], ChangKyoo Yoo[e], Hongbin Liu[a,c,d,*]

a Jiangsu Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, Nanjing Forestry University, Nanjing 210037, China

b College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

c Guangxi Key Laboratory of Clean Pulp & Papermaking and Pollution Control, College of Light Industry and Food Engineering, Guangxi University, Nanning 530004, China

d Laboratory for Comprehensive Utilization of Paper Waste of Shandong Province, Shandong Huatai Paper Co. Ltd., Dongying 257335, China

e Department of Environmental Science and Engineering, College of Engineering, Kyung Hee University, Yongin 446701, Republic of Korea

Corresponding Author:

*H.L. Tel.: +86-25-85427620; Fax: +86-25-85428793;

E-mail: hongbinliu@njfu.edu.cn

Frontiers of Environmental Science & Engineering

June 10, 2023

Table S1. The hyperparameters details of ANN and VAER

Models	Number of layers	Number of neurons in each layer	Dropout	Learning rate	Epochs	Batch size
ANN	6	(32,64,96,96,64,1)	0.2	0.001	800	128
VAER	8	(160,112,56,28) (28,56,112,160)	0.2	0.001	1000	64

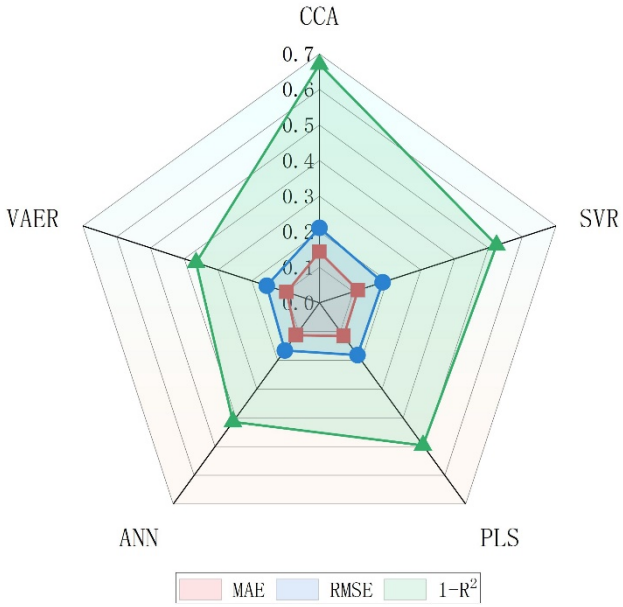

Figure S1. The radar plot of different single models on the test set.

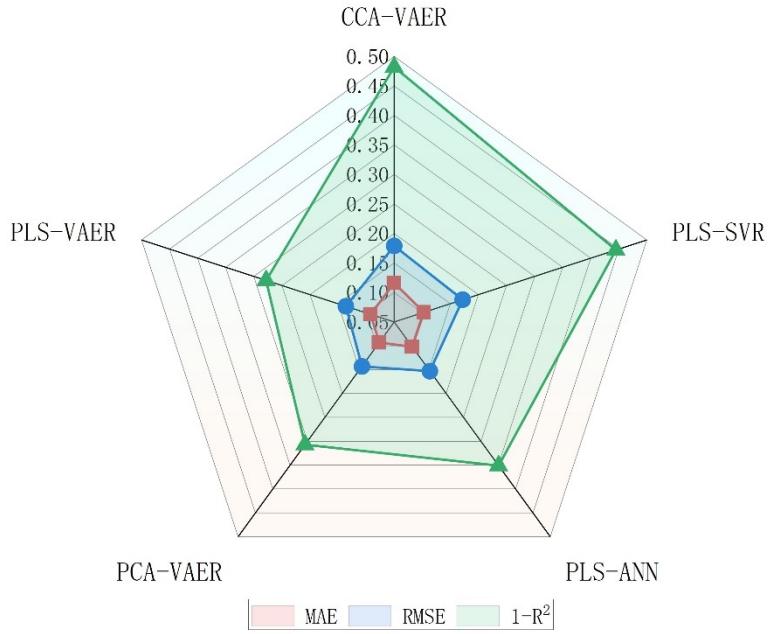

Figure S2. The radar plot of different hybrid models on the test set.

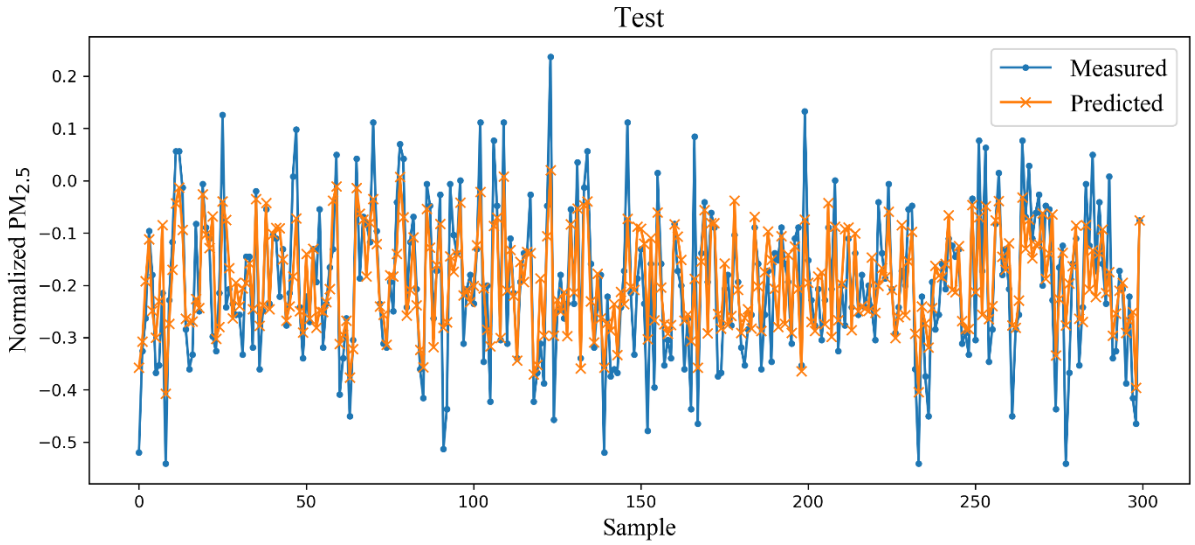

Figure S3. The prediction result of PLS-VAER model on the test set.

Supplementary experiment

1. Wastewater treatment process data

We used the wastewater data collected from an urban wastewater treatment plant in June, 1993 to verify the universality of the PLS-VAER model proposed in this paper. The data consists of 527

samples with 38 variables. The statistical information of the dataset is shown in Table S2. For all the models in this experiment, the concentration of effluent biological oxygen demand is the output and the remaining are used as inputs. After the preprocessing step, 379 samples remained. 303 of the samples are utilized as training sets and 76 as test sets.

Table S2. Statistical information of wastewater treatment process data

Variables	Mean	Standard deviation	Min	Max
input flow to plant	37226.56	6571.46	10000	60081
input Zinc to plant	2.36	2.74	0.1	33.5
input pH to plant	7.81	0.24	6.9	8.7
input biological demand of oxygen to plant	188.71	60.69	31	438
input chemical demand of oxygen to plant	406.89	119.67	81	941
input suspended solids to plant	227.44	135.81	98	2008
input volatile suspended solids to plant	61.39	12.28	13.2	85.0
input sediments to plant	4.59	2.67	0.4	36
input conductivity to plant	1478.62	394.89	651	3230
input pH to primary settler	7.83	0.22	7.3	8.5
input biological demand of oxygen to primary settler	206.20	71.92	32	517
input suspended solids to primary settler	253.95	147.45	104	1692
input volatile suspended solids to primary settler	60.37	12.26	7.1	93.5
input sediments to primary settler	5.03	3.27	1.0	46.0

input conductivity to primary settler	1496.03	402.58	646	3170
input pH to secondary settle r	7.81	0.19	7.1	8.4
input biological demand of oxygen to secondary settler	122.34	36.02	26	285
input chemical demand of o xygen to secondary settler	274.04	73.48	80	511
input suspended solids to se condary settler	94.22	23.94	49	244
input volatile suspended soli ds to secondary settler	72.96	10.34	20.2	100
input sediments to secondary settler	0.41	0.37	0.0	3.5
input conductivity to second ary settler	1490.56	399.99	85	3690
output pH	7.70	0.18	7.0	9.7
output biological demand of oxygen	19.98	17.20	3	320
output chemical demand of oxygen	87.29	38.35	9	350
output suspended solids	22.23	16.25	6	238
output volatile suspended sol ids	80.15	9.00	29.2	100
output sediments	0.03	0.19	0.0	3.5
output conductivity	1494.81	387.53	683	3950
performance input biological demand of oxygen in prim ary settler	39.08	13.89	0.6	79.1
performance input suspended solids to primary settler	58.51	12.75	5.3	96.1
performance input sediments to primary settler	90.55	8.71	7.7	100
performance input biological demand of oxygen to secon dary settler	83.44	8.4	8.2	94.7

performance input chemical demand of oxygen to second ary settler	67.67	11.61	1.4	96.8
global performance input bio logical demand of oxygen	89.01	6.78	19.6	97
global performance input che mical demand of oxygen	77.85	8.67	19.2	98.1
global performance input sus pended solids	88.96	8.15	10.3	99.4
global performance input sed iments	99.08	4.32	36.4	100

2. Results

Figure S4 shows the prediction result of the proposed PLS-VAER model on the test set.

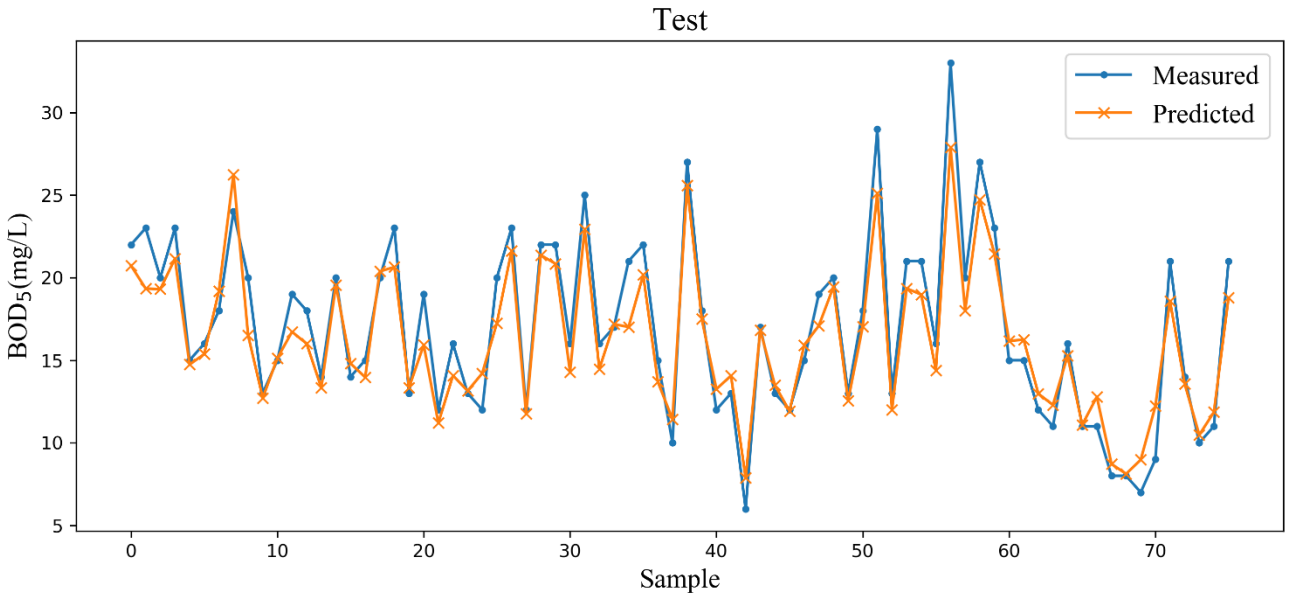

Figure S4. The prediction result of PLS-VAER on the test set.

Table S3 shows the prediction results of the proposed PLS-VAER model and other comparison models. It can be seen from the Table S3 that for this set of data sets, the accuracy of the proposed PLS-VAER model is still the highest among all test models, which shows that the proposed method has certain universality to a certain extent.

Table S3. Comparison of the prediction results of all models

Methods	Training Set			Test Set		
	MAE	RMSE	R^2	MAE	RMSE	R^2
PLS	2.286	3.029	0.812	2.337	2.945	0.695
SVR	2.335	3.890	0.690	2.168	2.843	0.716
ANN	1.660	2.270	0.895	2.146	2.682	0.747
VAER	1.592	2.409	0.881	1.590	2.250	0.822
PLS-SVR	2.096	4.125	0.652	1.682	2.398	0.798
PLS-ANN	1.388	1.983	0.920	1.708	2.094	0.846
PCA-VAER	1.093	1.563	0.942	1.673	1.934	0.853
PLS-VAER	0.897	1.316	0.965	**1.368**	**1.708**	**0.898**