Т. 90, № 4 ЖУРНАЛ ПРИКЛАДНОЙ СПЕКТРОСКОПИИ ИЮЛЬ — АВГУСТ 2023
V. 90, N 4 JOURNAL OF APPLIED SPECTROSCOPY JULY — AUGUST 2023

FORECAST OF OIL CONTENT IN OILFIELD WASTEWATER BY PLS AND CNN BASED ON UV TRANSMITTANCE SPECTRUM AND TURBIDITY**

Qiushi Wang [1,2], Haolin Li [1], Hanbing Qi [1,2], Haiqian Zhao [3*], Huaizhi Li [1], Xiaoxue Zhang [1]

*[1] School of Architecture and Civil Engineering, Northeast Petroleum University,
Fazhan Lu Street, Daqing, China*
*[2] Heilongjiang Key Laboratory of Petroleum and Petrochemical Multiphase Treatment
and Pollution Prevention, Daqing, China*
*[3] School of Mechanical Science and Engineering, Northeast Petroleum University,
Daqing, China; e-mail: dqzhaohaiqian@163.com*

Oil content plays an important role in oilfield wastewater treatment. To investigate the forecast of oil content by UV spectrophotometry, samples of oilfield wastewater are collected, and their UV transmittance and turbidity are measured. Partial least squares (PLS) and convolutional neural networks (CNN) based on a dataset of UV transmittance spectra are used for quantitative analysis in this work. The correlation coefficient between the oil content and turbidity of oilfield wastewater is 0.924, which shows a high positive linear correlation between the oil content and turbidity. Turbidity is added to the dataset to investigate its influence on the accuracy of prediction. The results show that the accuracy of models built by transmittance and turbidity is higher than that of models built by transmittance only, which is confirmed for both PLS and CNN. With the same dataset composition, the PLS and CNN models are nearly accurate, but the CNN performs slightly better overall. This work laid the foundation for the prediction of oil content in oilfield wastewater based on UV spectrophotometry and the further implementation of online detection.

Keywords: *UV transmittance spectrum, convolutional neural networks, oilfield wastewater, oil content, turbidity.*

ОПРЕДЕЛЕНИЕ СОДЕРЖАНИЯ НЕФТИ В СТОЧНЫХ ВОДАХ НЕФТЕПРОМЫСЛОВ МЕТОДОМ ЧАСТИЧНЫХ НАИМЕНЬШИХ КВАДРАТОВ И СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ НА ОСНОВЕ СПЕКТРА ПРОПУСКАНИЯ УФ-ИЗЛУЧЕНИЯ И МУТНОСТИ

Q. Wang [1,2], Haolin Li [1], H. Qi [1,2], H. Zhao [3*], Huaizhi Li [1], X. Zhang [1]

УДК 543.42.062:665.6

*[1] Школа архитектуры и строительства Северо-Восточного нефтяного университета,
Дацин, Китай*
*[2] Хэйлунцзянская лаборатория многофазной обработки нефти и нефтехимии
и предотвращение загрязнения, Дацин, Китай*
*[3] Школа машиностроения и инженерии Северо-Восточного нефтяного университета,
Дацин, Китай; e-mail: dqzhaohaiqian@163.com*

(Поступила 13 июля 2022)

Для исследования содержания нефти с помощью УФ-спектрофотометрии отобраны пробы сточных вод нефтепромысла и измерены их УФ-пропускание и мутность. Для количественного анализа использованы метод частичных наименьших квадратов (PLS) и сверточные нейронные сети (CNN), основанные на наборе данных спектров УФ-пропускания. Коэффициент корреляции между содержанием нефти и мутностью сточных вод нефтепромысла 0.924 свидетельствует о линейной

** Full text is published in JAS V. 90, No. 4 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 90, No. 4 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

зависимости между содержанием нефти и мутностью. Мутность добавляется в набор данных для исследования ее влияния на точность прогноза. Показано, что точность моделей, построенных по коэффициенту пропускания и мутности, выше, чем у моделей, построенных только по коэффициенту пропускания, что подтверждается как для PLS, так и для CNN. При одинаковом составе набора данных модели PLS и CNN почти точны, но в целом CNN работает немного лучше. Положено начало прогнозированию нефтесодержания в сточных водах нефтепромыслов на основе УФ-спектрофотометрии и дальнейшему внедрению оперативного детектирования.

Ключевые слова: *спектр пропускания УФ-излучения, сверточные нейронные сети, сточные воды нефтепромыслов, нефтесодержание, мутность.*

Introduction. Oilfield wastewater is a complex mixture containing oil, organic and inorganic matter and other compounds dissolved in water that ranges from fresh to brine [1], in which oil content detection is essential for oilfield wastewater treatment. Currently, common optical measurement methods include fluorescence spectrophotometry [2, 3], infrared spectrophotometry [4], and ultraviolet (UV) spectrophotometry [5, 6]. UV spectrophotometry has been used in wastewater quality detection based on partial least squares (PLS) and artificial neural networks (ANNs) [7–9]; however, there are few studies on the quantitative analysis of oil content in oilfield wastewater based on UV spectrophotometry. In this work, samples of oilfield wastewater are collected for the quantitative analysis of oil content prediction. PLS and convolutional neural network (CNN) models are built based on their UV transmittance spectra to compare their prediction accuracy. In recent years, many studies have shown that turbidity has an inevitable influence on the predicted accuracy based on UV spectrophotometry [10, 11]. Y. Hu et al. [12] proposed a method that deduced the turbidity component from surrogate parameters based on the proportion of four parameters in a formazine turbid solution to eliminate the impact of turbidity in water contaminant analysis. To investigate the influence of adding turbidity into the training set, PLS and CNN models based on the UV transmittance spectrum and turbidity are established.

Materials and method. A total of 96 samples in this work were of oilfield wastewater originating from the two oil extraction plants in Daqing. The sample transmittance of wavelengths ranging 190–900 nm was measured by a UV-Vis spectrophotometer (TU1900, Purkinje General Instrument, Beijing), and the turbidity of the samples was measured by a turbidity meter (JC-WGZ-1A, Juchuang Environmental Protection Group, Qingdao). The optical path was 10 mm, and the environmental temperature was 293.15 K. The dataset of 96 samples was collected and separated into a training set and a test set at a 5:1 ratio. All transmittance of the samples was preprocessed by normalization to reduce noise. The original transmittance spectrum of the test set and statistical results are shown in Fig. 1 and Table 1, respectively.

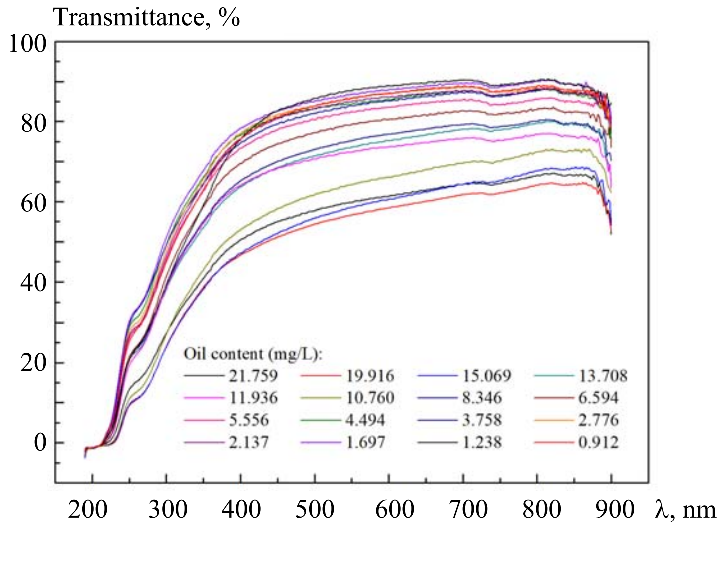

Fig. 1. Original transmittance spectrum of the test set.

TABLE 1. Statistical Results of the Oil Content in Samples

Dataset	Maximum	Minimum	Mean	Standard deviation
Train	31.603	0.764	7.701	6.580
Test	21.759	0.912	8.166	6.482

Partial least squares (PLS), as a method commonly used to build an analysis model of spectroscopy [13–15], is used to investigate the relationship between two matrices. PLS is a common linear model that creates a linear regression by projecting both the predicted and real variable values into a new space [16]. The procedure is as follows. A dataset named $\{X,Y\}$ consists of two matrices, where X is the process variable data matrix, and Y is the corresponding dependent variable vector:

$$X = TP^T + E,$$
$$Y = UQ^T + F,$$
$$U = TB^T,$$

where T, P, and E are the score matrix, loading matrix and residual matrix of X, while U, Q, and F are the score matrix, loading matrix and residual matrix of Y, respectively. B is the regression coefficient matrix of U and T:

$$Y_P = X_P B,$$

where X_P and Y_P are the transmittance matrices of the validation set and predicted oil content, respectively.

Artificial neural networks (ANNs) have been one of the most popular methods employed to address multiple regression problems. As a kind of ANN, CNN has been widely applied in the field of image identification, and it has begun to be used in spectrum analysis in recent years [17, 18]. Compared with image identification by two-dimensional (2D) CNN, the sample set in spectrum analysis is a matrix of data that should be input into a one-dimensional (1D) CNN. The CNN architecture used in this work is composed of an input layer, convolution layer, batch normalization layer, flattening layer, fully connected layer and output layer.

The input layer is set to accept UV transmittance data of one sample in the entire range of measured wavelengths. Then, the features of the input data are collected by the convolution layer. Compared with the dense layer, which studies global data, the convolution layer studies data in a sliding window on the input data, which makes the model utilize data efficiently and perform better with a smaller sample size. To reduce the vanishing gradient, exploding gradient and overfitting, a batch normalization layer is added after the convolution layer. The output data cannot be input into a fully connected layer, so a flattening layer is necessary to compress data into a 1D array. The fully connected layer has the function of feature classification. Then, the result is output according to the output layer.

At the beginning of model design, an overfitting CNN architecture of large size is set for the best training accuracy. The validation accuracy of the model based on this architecture is less than the training accuracy because of overfitting. Then, the hyperparameters in the architecture are optimized to correct overfitting. Finally, the best model architecture with the highest validation accuracy is completed.

When the sample size is insufficient, the accuracy of different models varies widely due to the dissimilar proportion of dividing the dataset into a training set and a validation set, so the performance of the model built by the invariant training set is not representative. As shown in Fig. 2a, the distribution of data used in this paper is not uniform, with oil contents between 16–27 mg/L accounting for 8.75%, and 0.9–1 mg/L accounting for 18.75%. Therefore, conventional K-fold validation is not suitable for this work.

Bootstrap sampling is more efficient. The original dataset was divided into nine value intervals. The data in the test set were randomly selected from each interval according to the ratio of the number of samples in the interval to the dataset. As shown in Fig. 2b, the test set has a similar proportion distribution as the original dataset, and the same is true for the training set. In this work, 16 sample data points were taken from the dataset as the test set, and the remaining 80 sample data points were used as the training set. Bootstrap sampling was used in the training of CNN models. The training set was again divided into a training set and a validation set by bootstrap sampling to build the model. The process was repeated ten times, and the model was evaluated by the average accuracy for all. The optimized architecture and hyperparameters were used in the model trained by all 80 samples and then tested by the test set.

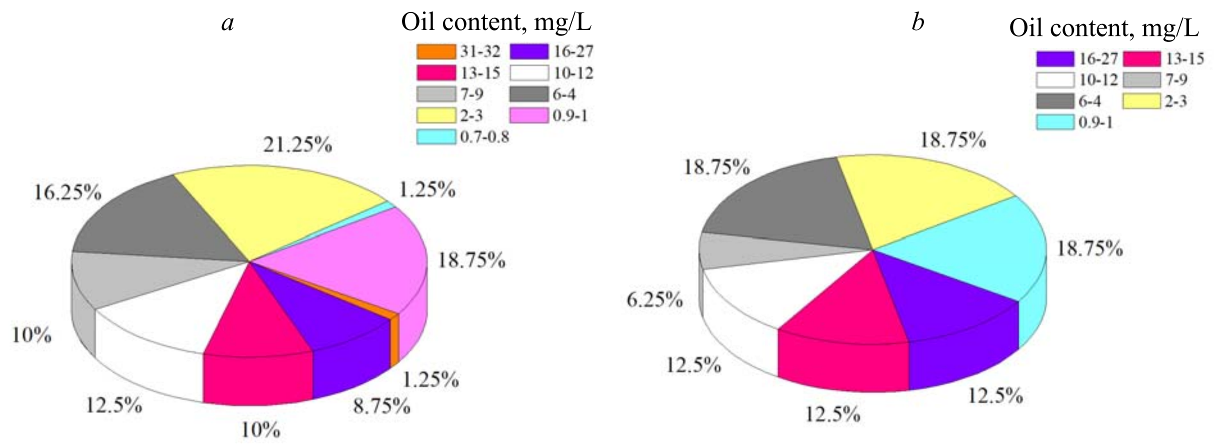

Fig. 2. Distribution of the (a) original dataset and (b) test set.

Results and discussion. As shown in Fig. 3, the Pearson correlation coefficient between the oil content and turbidity of oilfield wastewater is 0.924, which indicates that there is a high positive linear correlation between the oil content and turbidity in oilfield wastewater. Therefore, it can be assumed that turbidity could be combined with transmittance to build a model predicting the oil content of oilfield wastewater for better accuracy than the model built by transmittance only.

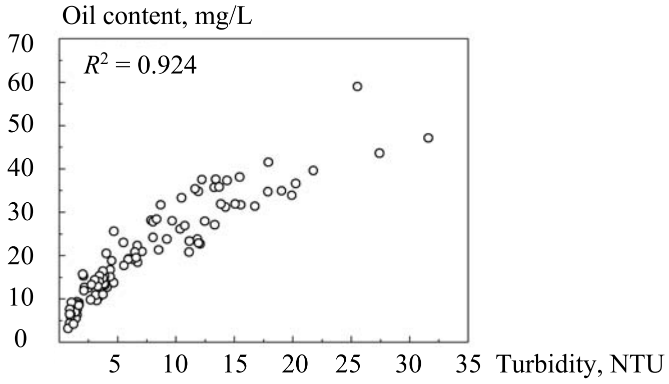

Fig. 3. Correlation between oil content and turbidity in oilfield wastewater.

Two PLS models named PLS-T and PLS-T-T were established to test this hypothesis. The models were built based on the dataset of transmittance and the combination of turbidity and transmittance. The predicted results of the test set by the two models are shown in Fig. 4.

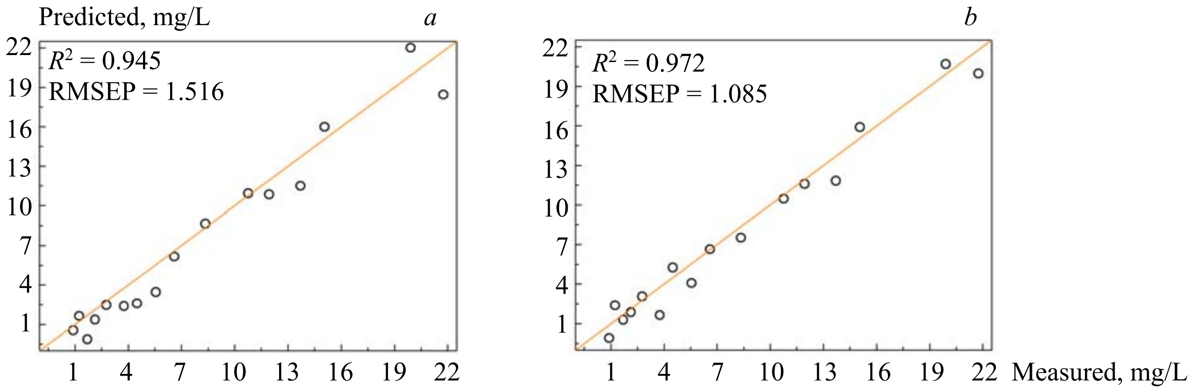

Fig. 4. Result of (a) PLS-T and (b) PLS-T-T.

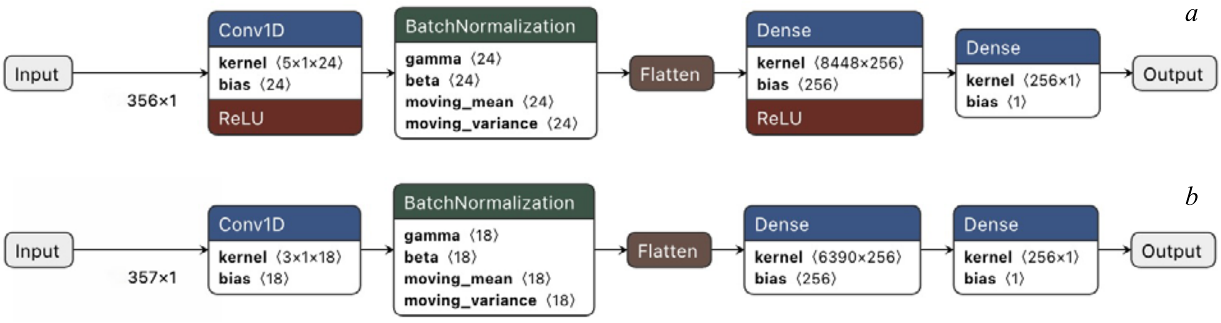

Fig. 5. Architecture of CNN-T (a) and CNN-T-T (b).

The predictive R^2 and RMSEP were 0.945 and 1.516 mg/L for PLS-T, while those of PLS-T-T were 0.972 and 1.085 mg/L, respectively. Obviously, the result by PLS-T-T had higher R^2 and lower RMSEP, which means that PLS-T-T had better accuracy than PLS-T. Therefore, PLS accuracy is effectively improved by adding turbidity to the dataset.

Two CNN models named CNN-T and CNN-T-T, based on transmittance and the combination of turbidity and transmittance, respectively, were established to evaluate the accuracy of predicting oil content in oilfield wastewater by a CNN. The architecture was optimized in the bootstrap sampling of the transmittance training set to overcome overfitting. The best CNN architecture optimized to obtain the minimum predicted MAE is summarized in Table 2 and Fig. 5.

TABLE 2. Hyperparameters of the CNN Models

Model	Activation	Learning rate	Batch size	Epochs
CNN-T	Relu	3.00E-05	356	3489
CNN-T-T	Linear	3.00E-05	357	1996

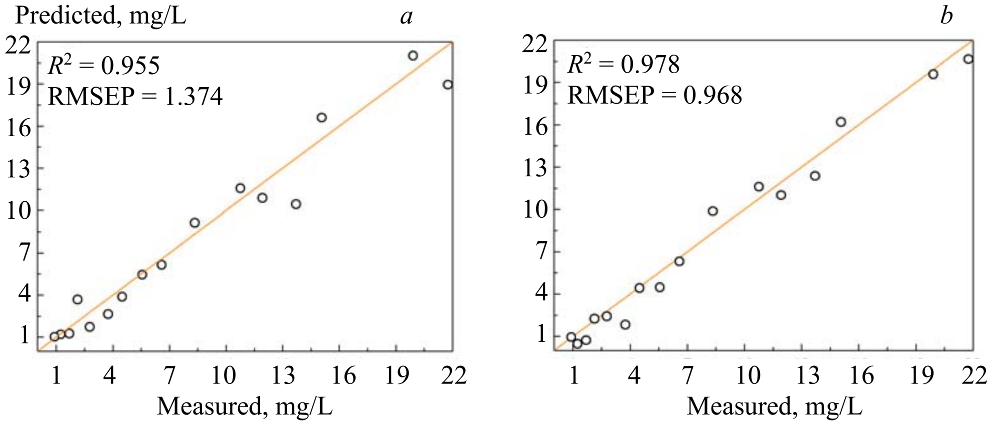

Fig. 6. Result of (a) CNN-T and (b) CNN-T-T.

The CNN architecture is affected by many factors in which the sample type and sample size have a great influence; thus, two different architectures were used in CNN-T and CNN-T-T. Initially, a deep network was set to overfit the model, and then the filter size and layer size were decreased for better validation accuracy; furthermore, the max pooling layer, which decreases computation and extracts features, was not used in this work because the dataset in this work is insufficient compared with the image recognition field and valuable features can be lost when sifted out by the max pooling layer. Meanwhile, this is also the reason why the dropout layer was not used. The model without the max pooling layer and dropout layer had better accuracy during model training, and the MAE was 0.1–0.2 lower. The results are shown in Fig. 6. The R^2 and RMSEP of CNN-T were 0.955 and 1.374, respectively, while those of CNN-T-T were 0.978 and 0.968, respectively. Apparently, the CNN-T-T had better performance than CNN-T overall. The distribution of predicted oil content in the range of high oil content (higher than 12 mg/L) in CNN-T is spread out compared to that of CNN-T-T; moreover, the result by CNN-T-T has a lower error than that of CNN-T, which indicates that in the same sample size, the CNN model based on the combination of transmittance and turbidity made better predictions than that based on transmittance.

When transmittance is used as the dataset to establish the model of predicting oil content in oilfield wastewater, comparing the PLS and CNN, R^2 of CNN is 0.01 higher than that of PLS, while RMSEP is 0.0142 lower. When the dataset contains transmittance and turbidity data, R^2 of the CNN is 0.006 higher than that of the PLS, while RMSEP is 0.117 lower; therefore, the CNN model had higher accuracy than PLS under the same conditions. Overall, for both CNN and PLS, the model based on the combination of transmittance and turbidity performs better than that based on transmittance only.

The CNN model has more tunable parameters and randomness than PLS; therefore, for the CNN model presented in this work, the architecture was used to repeat the building model 30 times to evaluate the robustness of the architecture. The average R^2 and RMSEP values were calculated separately for comparison with PLS. The results are shown in Figs. 7 and 8.

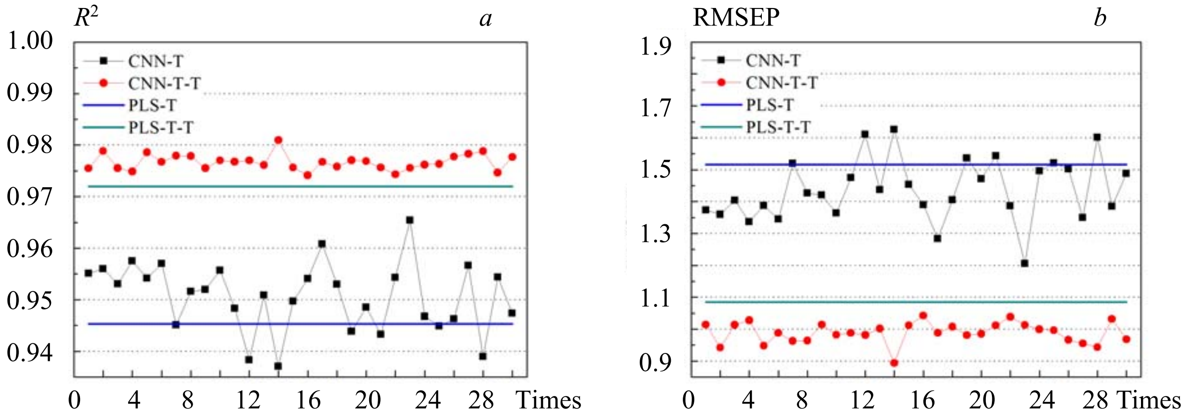

Fig. 7. R^2 (a) and RMSEP (b) of 30 replications.

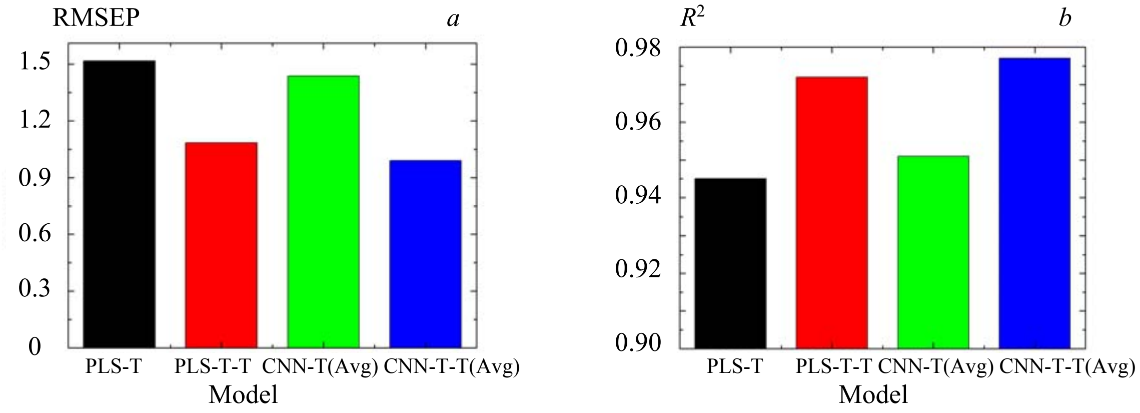

Fig. 8. Average result of 30 replications.

As shown in Fig. 7, it is obvious that for CNN-T-T, most R^2 values were higher than those of the other models, and the corresponding RMSEP was the lowest. All R^2 and RMSEP values of CNN-T-T were in the ranges of 0.974–0.981 and 0.894–1.043, respectively. For CNN-T, the ranges were 0.937–0.966 and 1.205–1.627, respectively. The results show that the architecture of CNN-T-T has better robustness than that of CNN-T. As shown in Fig. 8, the average R^2 of CNN-T-T was 0.026 higher than that of CNN-T, and RMSEP was 0.448 lower, while PLS-T-T was also more precise than PLS-T because of the better R^2 and RMSEP. Therefore, the PLS and CNN models can be made more accurate by adding turbidity to the dataset.

For the transmittance dataset, the R^2 (average) of CNN-T was 0.006 higher than that of PLS-T, and RMSEP was 0.079 lower, which indicates that CNN-T had less error than PLS-T, while their R^2 values were almost the same. When turbidity was added to the dataset, the R^2 (average) and RMSEP values of CNN-T-T were 0.005 higher and 0.096 lower than those of PLS, respectively. CNN was more accurate than PLS overall.

In summary, there is no dramatic gap in accuracy between PLS and CNN, although CNN performs better than PLS in this study. Nevertheless, CNN architecture with more hyperparameters and long training time is much more difficult to train than PLS under the same sample size. Theoretically, the accuracy of CNN can be improved by increasing the sample size [19], but the oil content of samples requires too much time to measure in the laboratory. Moreover, the models built based on the same CNN architecture perform differently because of the randomness of ANNs; therefore, PLS is more suitable for sample sets with small sizes. On the other hand, for a fair comparison, the transmittance of all wavelengths was used in both PLS and CNN. However, the accuracy of PLS can be improved by selecting wavelengths with useful information such as interval partial least squares (iPLS) [20] and MWPLS [21]; thus, it can be assumed that the CNN can also be improved in the same way.

Conclusions. The UV transmittance spectrum of oil content in oilfield wastewater was measured. The correlation coefficient between oil content and turbidity was calculated. PLS and ANN models based on the transmittance dataset to predict the oil content of oilfield wastewater were compared. Then, turbidity was added to the dataset to rebuild the models. The following conclusions can be drawn from the preceding study. Firstly, there is a high positive linear correlation between oil content and turbidity in oilfield wastewater, with a correlation coefficient of 0.924. Secondly, for the UV transmittance dataset, the average R^2 and RMSEP values of the CNN models built 30 times were better than the accuracy of the PLS. However, the results were not stable, with R^2 values in the range of 0.937–0.966 and RMSEP values in the range of 1.205–1.627. Thus, for the transmittance and turbidity datasets, CNN had dramatically better accuracy than PLS with good robustness. Lastly, PLS and CNN models based on the combination of UV transmittance and turbidity had higher R^2 and lower RMSEP than models built using transmittance alone. Therefore, it is useful to add turbidity into the dataset to improve the accuracy of the PLS and CNN models in predicting the oil content in oilfield wastewater.

Acknowledgements. This work was supported by the Youth Innovative Talents Training Plan of General Undergraduate University in Heilongjiang Province (UNPYSCT-2020148), the Open project of MOE Key Laboratory for Enhanced Oil and Gas Recovery (No. NEPU-EOR-2022-06), and the Postdoctoral support project of Heilongjiang Province (LBH-Q21084).

REFERENCES

1. F. Al Jabri, L. Muruganandam, D. A. Aljuboury, *Global NEST J.*, **21**, No. 2, 204–210 (2019).

2. Xu Shirong, Duan Ming, Zhang Jian, *Chem. Eng. Oil and Gas*, **38**, No. 3, 258–261 (2009).

3. C. Y. Wang, H. H. Jiang, J. W. Gao, J. L. Zhang, R. E. Zheng, *Spectrosc. Spectr. Anal.*, **26**, No. 6, 1080–1083 (2006).

4. Yang Haiya, Wu Liangzhuan, Yu Yuan, Zhi Jinfang, *Chin. J. Spectrosc. Lab.*, **28**, No. 6, 2770–2773 (2011).

5. P. D. Wentzell, D. T. Andrews, J. M. Walsh, J. M. Cooley, et al., *Can. J. Chem.*, **77**, No. 3, 391–400 (1999).

6. X. Bian, S. Li, L. Lin, X. Tan, Q. Fan, M. Li, *Anal. Chim. Acta*, **925**, 16–22 (2016), doi: 10.1016/j.aca.2016.04.029.

7. Xiaojun Tang, Angxin Tong, Feng Zhang, Bin Wang, *Sains Malaysiana*, **49**, No. 8, 1773–1785 (2020).

8. P. Li, J. Qu, Y.He, Z. Bo, M. Pei, *RSC Adv.*, **10**, No. 35, 20691–20700 (2020), doi: 10.1039/c9ra10732k.

9. W. S. Jia, H. Z. Zhang, J. Ma, G. Liang, J. H. Wang, X. Liu, *Spectrosc. Spectr. Anal.*, **40**, No. 9, 2981–2988 (2020).

10. X. W. Chen, G. F. Yin, N. J. Zhao, T. T. Gan, R. F. Yang, W. Zhu, J. G. Liu, W. Q. Liu, *Spectrosc. Spectr. Anal.*, **39**, No. 9, 2912–2916 (2019).

11. Wu Decao, Wei Biao, Tang Ge, Feng Peng, Tang Yuan, Liu Juan, Xiong Shuangfei, *Acta Opt. Sin.*, **37**, No. 2, 0230007 (2017).

12. Y. Hu, X. Wang, *Sensors and Actuators B: Chem.*, **239**, 718–726 (2017), doi: 10.1016/j.snb.2016.08.072.

13. E. Carré, J. Pérot, V. Jauzein, L. Lin, M. Lopez-Ferber, *Water Sci. Technol.*, **76**, No. 3, 633–641 (2017), doi: 10.2166/wst.2017.096.

14. B. Chen, H. Wu, S. F. Y. Li, *Talanta*, **120**, 325–330 (2014), doi: 10.1016/j.talanta.2013.12.0.

15. X. Liu, L. Wang, *Water Sci. Technol.*, **71**, No. 10, 1444–1450 (2015), doi: 10.2166/wst.2015.110.

16. J. C. Cancilla, R. Aroca-Santos, K. Wierzchoś, J. S. Torrecilla, *Chemom. Intell. Lab. Systems*, **156**, 102–107 (2016), doi: 10.1016/j.chemolab.2016.05.

17. Y. Chen, L. Song, Y. Liu, L. Yang, D. Li, *Appl. Sci.*, **10**, No. 17, 5776 (2020), doi: 10.3390/app10175776.

18. G. Puertas, M. Vázquez, *J. Food Comp. Anal.*, **86**, 103350 (2020), doi: 10.1016/j.jfca.2019.10335.

19. B. Wei, K. Hao, X. Tang, Y. Ding, *Textile Res. J.*, 004051751881365 (2018), doi: 10.1177/0040517518813656.

20. L. Norgaard, A. Saudland, J. Wagner, et al., *Appl. Spectrosc.*, **54**, No. 3, 413–419 (2000).

21. F. Al Jabri, L. Muruganandamand, D. A. Aljuboury, *Global NEST J.*, **21**, No. 2, 204–210 (2019).