ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Hybrid Deep Learning-Based Air and Water Quality Prediction Model

Jungeun Yoon[1,2] Dasong Yu[1,2] and Youngjae Lee[2] *

1- Andong National University - Dept of ICT Convergence Engineering
1375 Gyeongdong-ro - Andong-si

2- ETRI - Dept of Regional Industry IT Conversion Team
1 Techno sunhwan-ro 10-gil, Dalseong-gun - Daegu

Abstract. This paper analyzes the impact of surrounding data on predicting air and water pollution levels by incorporating relevant features and examining their influence. By doing so, we can confirm the relationship between air and water pollution. A hybrid deep learning-based model is trained and various datasets and models are compared and analyzed. The proposed GCN-GRU model achieved the best results not only for $PM_{2.5}$ but also for Dissolved Oxygen. The hybrid model takes into account the spatial and temporal effects of data characteristics and provides more accurate environmental prediction information through correlation analysis.

1 Introduction

With the acceleration of industrialization and urbanization, the seriousness of environmental pollution has become apparent, leading to an increase in interest in the environment that affects our living environment and personal health. In particular, water and air are essential substances in our daily lives, and environmental pollution tends to increase as pollution worsens. Various methods and measures have been developed to address environmental pollution in different regions. Recently, various technologies for predicting air and water pollution, such as classic statistical techniques like AR and ARIMA[1], machine learning [2,3], or artificial intelligence such as deep learning [4,5,8], have been applied.

In addition to these framework technologies, in time-series data analysis, it is necessary to process data to select the optimal feature data that can be applied to the analysis. Increasing the amount of data unconditionally is not effective in improving prediction performance. Therefore, it is important to build a dataset that can improve prediction accuracy by adding data with related features. In this paper, we added related features to existing data and constructed a dataset by considering data from surrounding areas for air and water quality prediction. We proposed a hybrid model that combines deep learning models such as CNN or LSTM to learn prediction values considering the characteristics of the dataset.

*Authors would like to give thanks to the engineers from Waterkorea for their helps to set up datasets. This work was supported by the ICT R&D program of grant funded by Andong City-Hall [23AD1110].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2 Time Series Datasets

In this paper, air and water quality status is predicted using a machine learning model trained with data from 1 hour ago. Before training, the collected data is preprocessed to handle missing values using linear interpolation, which is represented by Equation (1)[6].

$$y = y_0 + (y_1 - y_0) \times \frac{x - x_0}{x_1 - x_0} \tag{1}$$

To account for varying data scales, normalization is performed using the Min-Max technique, which scales all values between 0 and 1 according to Equation (2)[7].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Time series data is then transformed into supervised learning input/target pairs and split into training, validation, and test sets at a ratio of 80:10:10. The training dataset is used to train the model, the validation dataset is used to optimize model parameters, and the test dataset is used to evaluate model performance.

2.1 Air Quality

To predict air quality pollution levels, a total of 13,104 data points were collected from January to June 2021, including hourly measurements of air quality, Meteorological, and other related data. The $PM_{2.5}$ value was predicted as a measure of air quality, with the other characteristics serving as input variables. The features included SO_2, NO_2, O_3, CO, Temperature, Wind Speed, Wind Direct, Humidity, vapor pressure, dew point temperature, local pressure, sea level pressure, visibility, and ground temperature. As shown in Fig. 1, the air quality and Meteorological data in the target(Andong-si) area were considered, along with the data from the surrounding areas, to improve the performance of the prediction model.

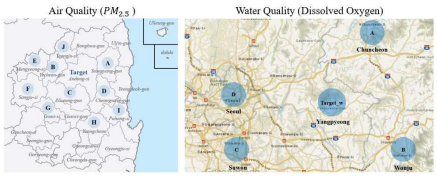

Fig. 1: Location of air quality and water quality data measured and collected

2.2 Water Quality

To predict water quality pollution levels, a total of 19,689 data points were collected from January 2017 to March 2019, including hourly measurements of air quality, Meteorological, and water quality data. The dissolved oxygen(DO) value

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

was predicted as a measure of water quality, with the other characteristics serving as input variables. The features included Water temperature, pH, Electrical Conductivity, DO, TP, TOC, Chlorophyll, Temperature, Wind Speed, Wind Direct, Humidity, vapor pressure, dew point temperature, local pressure, sea level pressure, visibility, ground temperature, $PM_{2.5}, PM_{10}, O_3$ and CO. As shown in Fig.1, the water quality data in the target-w(Yangpyeong) area were considered, along with the air quality and Meteorological data from the surrounding areas, to improve the performance of the prediction model.

3 Analysis Of Air And Water Quality

3.1 Models to forecasting

This paper considers data from the target region and surrounding areas. Therefore, a Convolutional Neural Networks(CNN) model that can reflect spatial characteristics in time series data is utilized to consider the spatial features of the data[8]. However, since CNN cannot reflect temporal characteristics, a hybrid model is constructed using Long Short-Term Memory Network(LSTM) or Gated Recurrent Unit(GRU) models [9,10] that can reflect a lot of past information in time series data. Fig. 2 shows the framework of the hybrid models implemented in this paper, which are the CNN-LSTM and CNN-GRU models.

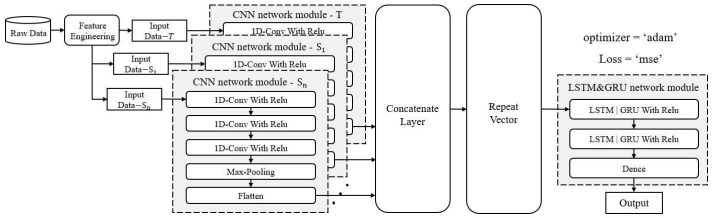

Fig. 2: The framework of the proposed hybrid prediction model.

In the case of CNN, it receives a lot of input data for each region, so it is computed in parallel using a multi-head technique to enable faster calculations. To extract features by reflecting spatial properties, it is composed of a three-step convolution layer, and the filter is set to 16, and the kernel size is set to 3. To prevent overfitting, the dropout layer is set to 0.2. Then, a two-step LSTM or GRU layer with 10 units is constructed to reflect temporal properties, and the final Dense layer is trained for the prediction value.

To better incorporate spatial characteristics beyond the previously proposed hybrid model, we apply Graph Convolutional Networks(GCN) in this study. GCN is a graph-based model that applies convolutional concepts from CNN and takes node features and adjacency matrices as input. Nodes represent each region, and the adjacency matrix is calculated based on the Euclidean distance between the nodes' latitude and longitude. Fig. 3 shows the implemented GCN-LSTM and GCN-GRU models in this study.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

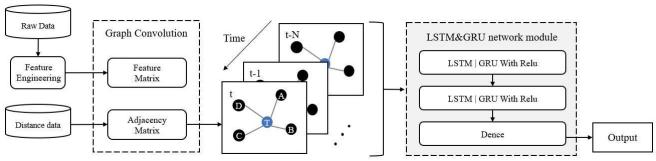

Fig. 3: Proposed GCN-LSTM&GRU Framework.

3.2 Performance Evaluation

To evaluate the performance of a prediction model, it is important to assess how well the predicted values match the actual values. Therefore, RMSE and MAE are used to calculate the error rate of the prediction model, and the closer these values are to 0, the better the performance of the prediction model. Additionally, R_2, which is used in regression models, is also considered as a performance metric, and the closer this value is to 1, the better the performance. These metrics are used to train the model and compare its performance.

3.3 Results: Performance of the prediction model

To compare the performance of different prediction models, two datasets were constructed in this study. The first dataset only considered the data from the prediction area, while the second dataset included the data from both the prediction area and its surrounding areas. For $PM_{2.5}$, up to eight areas were considered, and for DO, up to four areas were considered. For the latter, only Meteorological and air quality data were considered in the surrounding areas, while water quality data were not included.

	Methods							
	CNN-LSTM		CNN-GRU		GCN-LSTM		GCN-GRU	
Matrix	$PM_{2.5}$	DO	$PM_{2.5}$	DO	$PM_{2.5}$	DO	$PM_{2.5}$	DO
MAE	3.32	0.16	3.19	0.13	2.61	0.11	2.35	0.09
RMSE	5.24	0.22	5.10	0.18	3.41	0.14	3.17	0.11
R2	0.89	0.89	0.90	0.91	0.92	0.94	0.93	0.96

Table 1: The results of different models for $PM_{2.5}(\mu g/m_3)$ and DO(mg/L)

After training each model with these two datasets, we found that the models performed better when the surrounding areas were taken into account. The results are shown in Table 1, and both for $PM_{2.5}$ and DO, the GCN-GRU model showed the best performance.

Fig. 4 shows a 72-hour visualization of the prediction results of the GCN-GRU model. The left side of the graph, which is denoted by the black dotted line, represents the data used for training, while the right side shows the predicted values as the orange line and the observed values as the green line.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

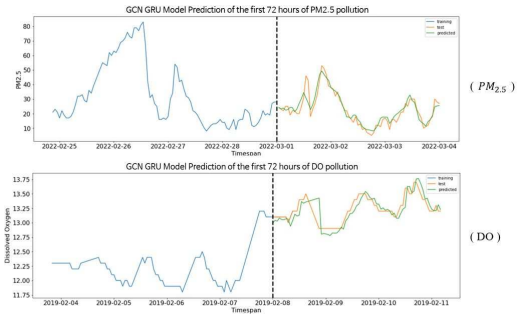

Fig. 4: The experimental results of the GCN-GRU model.

As shown in Fig. 4, the prediction performance of the model is high because the curves of the prediction data and the observation data are almost identical.

3.4 Performance Comparison

	$PM_{2.5}(\mu g/m_3)$		DO (mg/L)	
Matrix	GCN-GRU	Rakholia et al [11]	GCN-GRU	Huan et al [12]
MAE	2.35	3.25	0.09	0.30
RMSE	3.17	4.12	0.11	0.22
R2	0.93	-	0.96	-

Table 2: Comparison of models and other studies in this paper.

In this paper, we compare the performance of our implemented model with other studies, which can be observed in Table 2. For $PM_{2.5}$, Rakholia's study [11] implemented a 1D CNN-LSTM model to predict $PM_{2.5}$ values after 24 hours and evaluated prediction accuracy using MAE and RMSE. Our GCN-GRU model shows a lower performance of around 0.9 in MAE and 0.87 in RMSE, indicating a higher prediction accuracy. Regarding DO, Juan's study [12] implemented a GBDT-LSTM model and evaluated prediction accuracy using MAE and RMSE. Our GCN-GRU model shows a lower performance of around 0.21 in MAE and 0.11 in RMSE, indicating a higher prediction accuracy. In conclusion, our GCN-GRU model shows a high performance in predicting $PM_{2.5}$ and DO.

4 Conclusion

Recently, various approaches and measures have been developed to cope with environmental pollution, which has become severe and affects personal health. Machine learning, deep learning, and artificial intelligence are being applied to predict air and water pollution and improve accuracy. To improve prediction performance, this study constructed datasets as follows: Meteorological data was

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

added for air quality, and Meteorological and air quality data were considered for water quality. In addition, a method of considering data from surrounding areas was proposed. To reflect the characteristics of the proposed dataset, a hybrid prediction model, such as GCN-GRU, which considers spatial and temporal characteristics, was built and compared with previous studies.

The results showed that the GCN-GRU model trained with the dataset that considers the target and surrounding areas for both $PM_{2.5}$ for air quality and dissolved oxygen for water quality demonstrated superior performance compared to previous studies. Furthermore, for water quality, the prediction performance improved even by adding only the surrounding area's Meteorological and air quality data, indicating a correlation between the data. In the future, we will analyze the correlation between water and air quality, apply the trained model in real-time data collection environments, and improve the model to reflect the characteristics of real-time data.

References

[1] Q. An and M. Zhao, Time series analysis in the prediction of water quality, Advances in computer science research, vol. 76, pp. 51-54, 2017.

[2] K. Kim and J. Ahn, Machine learning predictions of chlorophyll-a in the Han river basin, Korea, Journal of Environmental Management 318, 2022

[3] H. Karimian, Qi Li, C. Wu, Y. Qi, Y. Mo, G. Chen, X. Zhang, and S. Sachdeva, Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations, Aerosol and Air Quality Research. Taiwan Asso. For Aerosol Research, vol.19, pp. 1400-410, 2019

[4] S. Du. T. Li, Y. Yanf and S. Horng, Deep air quality forecasting using hybrid deep learning framework, IEEE Trans. On knowledge and data engineering, vol. 33, issue. 6, pp. 271-350. 2019.

[5] K. Nagrecha, P. Muthukumar, E. Cocom, J. Holm, D. Comer, I. Burga and M. Pourhomayoun, Sensor based air pollution prediction using deep CNN-LSTM, Int. conf. on Computational Science and Computational Intelligence(CSCI), pp. 694-696, 2020.

[6] Y. Luo, and K. Lu, An online state of health estimation technique for lithium-ion battery using artificial neural network and linear interpolation, Journal of Energy Storage, 52, 2022.

[7] S. Kim, Y. Noh, Y. Kang, S. Park, J. Lee, and S. Chin, Hybrid datascaling method for fault classification of compressors, Measurement, 201, 2022.

[8] H. Hua, M. Liu, Y. Li, S. Deng, and O. Wang, An ensemble framework for short-term load forecasting based on parallel CNN and GRU with improved ResNet, Electric Power Systems Research, 216, 2023.

[9] Y. Yu, X. Si, C. Hu, and J. Zhang, A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures, Neural Computation, 31, July 2019

[10] VA, P. G., RV and KPS. DeepAirNet: Applying Recurrent Networks for Air Quality Prediction. Procedia Computer Science, 2018, 132: 1394-1403.

[11] R. Rakholia, Q. Le, K. Vu, B. Ho, and, R. S. Carbajo, AI-based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam, Urban Climate, 2022.

[12] J. Huan, H. Li, M. Li, and B. Chen, Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long shortterm memory network, Computers and Electronics inAgriculture, 175,June 2020.