2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 6-10, 2024. Sarawak, Malaysia

Multi-indicator Water Quality Prediction Using Multimodal Bottleneck Fusion and ITransformer with Attention

Jing Bi[1], Yibo Li[1], Xuan Zhang[1], Haitao Yuan[2], Ziqi Wang[1], Jia Zhang[3] and MengChu Zhou[4]

[1]College of Computer, Beijing University of Technology, Beijing 100124, China
[2]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
[3]Dept. of Computer Science in Lyle School of Engineering, Southern Methodist University, Dallas, TX, USA
[4]Dept. of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA

2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-1020-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMC54092.2024.10831495

***Abstract*—Water quality prediction methods forecast the future short or long-term trends of its changes, providing proactive advice for water pollution prevention and control. Existing water quality prediction methods only consider the historical data of single-type or multi-type water quality. However, meteorology and other factors also have a significant impact on water quality indicators. Therefore, only considering the historical data of water quality is not feasible. Unlike existing studies, this work proposes a hybrid water quality prediction model called CMI to solve the above problem. Before prediction, CMI incorporates a multimodal fusion mechanism of water quality time series and remote sensing images of meteorological rainfall. Moreover, CMI integrates the model of ConvNeXt V2 and a multimodal bottleneck transformer to extract image features for fusing the time series and images. Furthermore, it utilizes an emerging model of iTransformer to realize prediction with the fused features. Experimental results with real-life water quality time series and remotely sensed rainfall images demonstrate that CMI outperforms other state-of-the-art fusion algorithms, and the water quality prediction accuracy with fused meteorological data is 13% higher on average than that with only water quality time series.**

***Index Terms*—Water quality, multimodal fusion, time series prediction, multimodal bottleneck transformer, iTransformer**

I. INTRODUCTION

With the progress of civilization in human society and the enhancement of public awareness of environmental protection, the scientific usage and systematic protection of water resources have become an inevitable choice for the sustainable development of all countries in the world. Water quality prediction methods can obtain short or long-term water quality change trends in the future. Thus, it can guide water pollution prevention and provide technical support for water environmental control. Water quality predictions [1] are essentially a time series prediction problem, which refers to the prediction of changes in water quality indicators in the future period based on their values at historical time points. Water quality prediction research can be divided into mechanistic and data-driven models. Mechanistic models require a large number of parameters to be preset in advance and the training process is complex, requiring large computational resources and time costs. Data-driven models can be divided into statistical, machine learning, and deep learning methods. Machine learning is built upon statistical learning, and deep learning is a subfield of machine learning. They have achieved good results in the time series prediction in recent years. Compared with machine learning, which requires complex feature engineering, it can automatically learn patterns and trends in the time series data. Moreover, a neural network involves important parameters such as the number of hidden layers and the number of neurons.

Deep learning models is powerful to capture complex nonlinear patterns [2]. They can be summarized into two categories. **First**, recurrent neural networks (RNNs) [3], [4], *e.g.*, long short-term memory models [5], [6], are good at dealing with long sequences and effectively capture temporal dependencies in the time series. **Second**, convolutional neural networks (CNNs) [7] transform the time series data into a two-dimensional matrix and automatically extract its features through operations such as convolutional and pooling to realize the prediction. For example, temporal convolutional networks [8] solve problems of gradient vanishing and high computational complexity of traditional RNNs when dealing with the long series. In addition, RNNs and CNNs are often integrated with attention mechanisms to adaptively weight various parts of the input data, thus focusing more on the key information while reducing the influence of irrelevant information. For example, Transformer [9] treats the time steps of the input sequence as positional information, and designes the features of each time step as a vector and adopts the encoder-decoder framework for prediction. FEDformer [10] introduces a local attention mechanism and a reversible one to convert the time domain into the frequency domain, and it better captures local features in the time-series data and has higher computational efficiency.

However, there are many other factors affecting water quality indicators in the water environment, *e.g.*, meteorology, pollutants, and others. Thus, only considering the historical data of water quality is not sufficient [11] to make accurate

This work was supported in part by the National Natural Science Foundation of China under Grants 62073005 and 62173013, and the Beijing Natural Science Foundation under Grants 4232049 and L233005.

prediction. Other multimodal data, such as meteorological data, need to be jointly considered, Moreover, the fusion of data information from different modalities is imperative. Data fusion methods based on deep learning have been widely adopted to solve many complex environmental monitoring problems including water quality prediction. The work in [12] analyze the relationship between meteorological elements and wind power, and proposes a prediction method fusing wind speeds from multiple sources to predict the wind power generatio. In addition, multimodal fusion methods based on the attention mechanism have become popular. Multimodal bottleneck transformer (MBT) [13] introduces attention bottlenecks on top of the Transformer, forcing different modalities to share necessary information through a small number of bottlenecks. mPLUG [14] designs a cross-modal skip connection that allows visual modalities to skip cross-attention and directly performs self-attention, realizing efficient fusion. Liu *et al.* [15] design a loss function for medical image fusion in different dimensions and propose a multimodal feature fusion module to better preserve modality information.

Based on the aforementioned analysis, to improve the accuracy of water quality prediction with full utilization of meteorological data, a multi-indicator prediction model combining the ConvNeXt V2 [16], MBT, and ITransformer [17], called CMI for short, is proposed. The main contributions of this work are summarized as follows.

- Considering the influence of meteorological factors on water quality indicators, the multimodal fusion of water quality time series and remotely sensed rainfall images is innovatively proposed. The multimodal fusion-based prediction model called CMI is proposed with both time series and images as the input.
- CMI integrates ConvNeXt V2, MBT, and iTransformer to extract remotely sensed rainfall images features, realize multimodal fusion of time series and images, and predict future information with the fused information, respectively.
- Experimental results based on real-world water quality and remotely sensed rainfall images prove that the prediction accuracy with CMI is on average 13% higher than that with the non-fused one.

II. PROPOSED METHODOLOGY

This section presents the overall structure of CMI model. CMI is divided into three main components including data feature processing, multimodal data fusion, and prediction module. As shown in Fig. 1, time series are encoded using the embedding module. Moreover, remotely sensed rainfall images are dimension-aligned and feature-learned using the ConvNeXt V2 network. The fusion module includes the MBT for fusing water quality time series with remotely sensed rainfall images. Finally, the fused data is adopted as the input to the iTransformer for multi-indicator water quality prediction. Spewcifically, The input time series X_t is encoded into indicator tokens by the embedding block, and the images X_r are extracted by ConvNeXt V2 to generate uniform-format features y_t and y_r, which denote the tokens of the time series and remotely sensed rainfall images, respectively. y_t and y_r interact with each other through attention bottlenecks, whose intermediate results are further input into the fusion layer to produce the new information finally fed into iTransformer for the prediction.

A. Data Feature Processing

1) Embedding: The embedding structure of iTransformer is used to encode the water quality time series, which is different from the embedding in the traditional Transformer. Transformer embeds all the indicators of the same time node in the sequence into a time token. Thus, it does not differentiate between single and multiple indicators and pays more attention to the correlations between time nodes. When CMI deals with the multiple indicators, each indicator in the time series is embedded into the indicator token independently. Thus, it does not require the same time node, which enables clearer learning of the correlations among the indicators when the attention mechanism is utilized to describe interrelationships between tokens. Because the indicator tokens have positional logic within themselves, which can be implicitly stored in the neurons of the feed-forward neural network, CMI does not need the positional embedding in Transformer. The comparison of embeddings between Transformer and CMI is shown in Fig. 2.

2) ConvNeXt V2: ConNeXt V2 is adopted to extract features of remotely sensed rainfall images. It is built upon ConNeXt by designing a fully convolutional masked autoencoder framework, which consists of a sparse convolution-based ConvNeXt encoder and a lightweight ConvNeXt block decoder. Feature collapse occurs when training ConNeXt directly on masked inputs, and therefore, a global response normalization layer is added to address this issue to enhance feature competition among ConvNeXt block channels and promote feature diversity during the training.

B. Multimodal Fusion

CMI adopts the attention bottleneck fusion in MBT to fuse water quality time series data and remotely sensed rainfall images data. MBT is essentially Transformer applied to the multimodal case, and it introduces multiple new tokens $\mathbf{y}_{\mathrm{f}}=\left[y_{\mathrm{f}}^1, y_{\mathrm{f}}^2, \ldots, y_{\mathrm{f}}^B\right]$ as attention bottlenecks in the input data. B is the number of tokens in the attention bottleneck. The input sequence becomes $\mathbf{y}=[\mathbf{y}_{\mathrm{t}} \| \mathbf{y}_{\mathrm{f}} \| \mathbf{y}_{\mathrm{r}}]$. Different modalities can only share information and interact with each other through these bottleneck tokens. In this case, $\mathbf{y}_{\mathrm{t}}$ and $\mathbf{y}_{\mathrm{r}}$ can only exchange information through $\mathbf{y}_{\mathrm{f}}$. To reduce the computational complexity, the model requires that the information flow of each modality needs to be organized and condensed before passing through the bottleneck tokens, and the necessary information needs to be shared to ignore the redundant information. The number of attention bottleneck tokens needs to be restricted to be much smaller than the number of input data tokens. The bottleneck markers are updated separately according to different modes, one update each according to the time series and remotely sensed rainfall

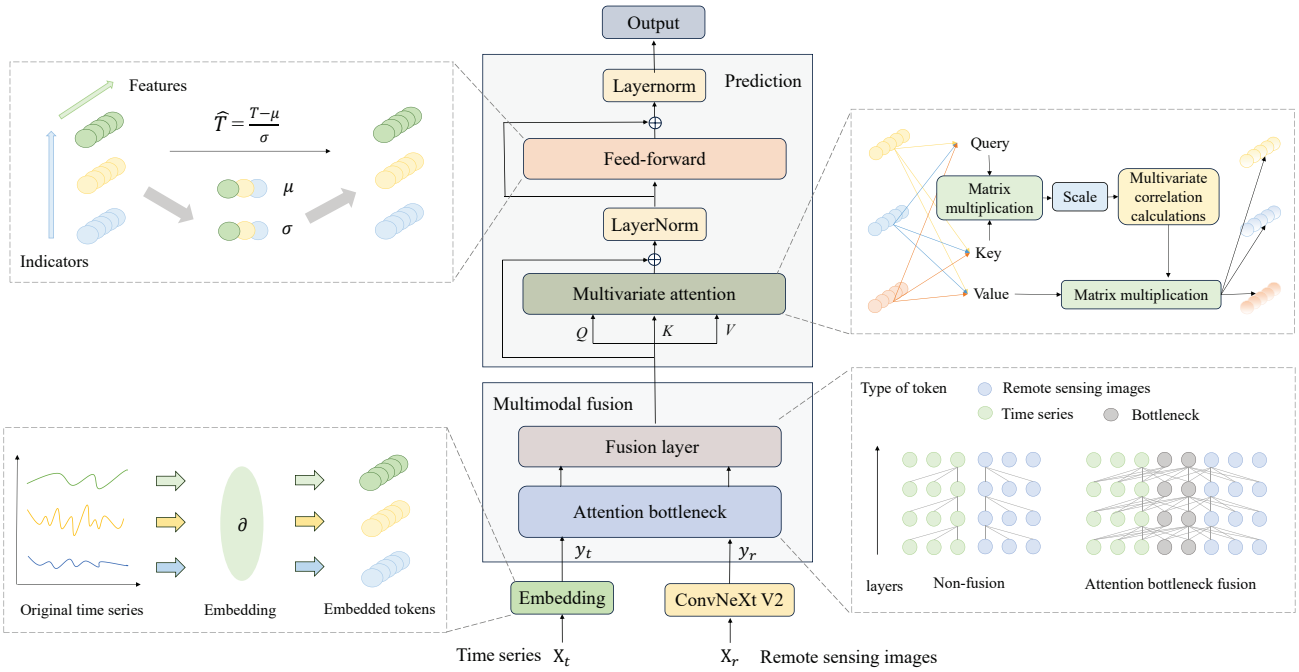

Fig. 1. Structure of CMI.

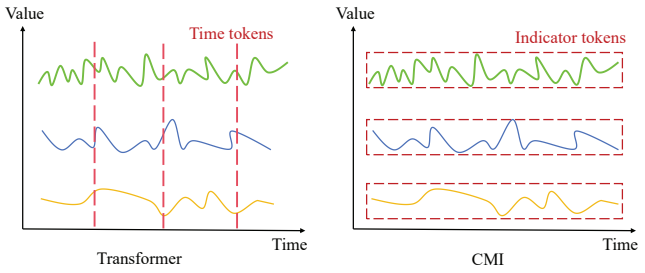

Fig. 2. Comparison of embeddings between Transformer and CMI.

images. Finally, the bottleneck markers of each mode are averaged to yield the final fusion markers. The process can be defined as:

$$\left[\mathbf{y}_{\mathrm{t}}^{l+1} \| \hat{\mathbf{y}}_{\mathrm{f}_t}^{l+1}\right] = \text{Transformer}\left(\left[\mathbf{y}_{\mathrm{t}}^{l} \| \mathbf{y}_{\mathrm{f}}^{l}\right]; \theta_{\mathrm{t}}\right) \tag{1}$$

$$\left[\mathbf{y}_{\mathrm{r}}^{l+1} \| \hat{\mathbf{y}}_{\mathrm{f}_r}^{l+1}\right] = \text{Transformer}\left(\left[\mathbf{y}_{\mathrm{r}}^{l} \| \mathbf{y}_{\mathrm{f}}^{l}\right]; \theta_{\mathrm{r}}\right) \tag{2}$$

$$\mathbf{y}_{\mathrm{f}}^{l+1} = \text{Avg}_{\mathrm{i}}\left(\hat{\mathbf{y}}_{\mathrm{f}_i}^{l+1}\right) \tag{3}$$

where $\mathbf{y}_{\mathrm{t}}^{l}$ denotes a vector of tokens of the time series in fusion layer l, $\mathbf{y}_{\mathrm{r}}^{l}$ denotes a vector of tokens of the remotely sensed rainfall images in fusion layer l, θ_{t} represents a parameter vector of the time series, and θ_{r} denotes a parameter vector of the remotely sensed rainfall images.

In terms of fusion location, a medium-term fusion strategy is employed, assuming that the number of fusion layers is in layer n, and each modality in the first $n-1$ layers learns its features with the self-attention mechanism, *i.e.*,

$$\mathbf{y}_{\mathrm{t}}^{l+1} = \text{Transformer}\left(\mathbf{y}_{\mathrm{t}}^{l}; \theta_{\mathrm{t}}\right) \tag{4}$$

$$\mathbf{y}_{\mathrm{r}}^{l+1} = \text{Transformer}\left(\mathbf{y}_{\mathrm{r}}^{l}; \theta_{\mathrm{r}}\right) \tag{5}$$

Then, $\mathbf{y}_{\mathrm{t}}^{l}$ and $\mathbf{y}_{\mathrm{r}}^{l}$ interact with each other through the attention bottlenecks in the fusion layer in a self-learning manner. $\mathbf{y}^{l}$ denotes a set of $\mathbf{y}_{\mathrm{t}}^{l}$ and $\mathbf{y}_{\mathrm{r}}^{l}$, which is given as:

$$\mathbf{y}^{l} = \left[\mathbf{y}_{\mathrm{t}}^{l} \| \mathbf{y}_{\mathrm{r}}^{l}\right] \tag{6}$$

$$\mathbf{y}^{l+1} = \text{Multimodal-Transformer}\left(\mathbf{y}^{l}; \theta_{\mathrm{t}}, \theta_{\mathrm{r}}\right) \tag{7}$$

C. Multi-indicator Water Quality Prediction

This work adopts iTransformer to implement multi-indicator water quality prediction. The structure is basically the same as the encoder of transformer, and its module fucntions have been changed because of different embedding methods and the changes of operands from time tokens to indicator tokens. In iTransformer, a feed forward neural network learns non-linear characteristics of each indicator token, which encodes a individual token and decodes the future representation. A normalization layer is used to normalize indicator tokens, which keeps different indicator variables in the same interval and reduces differences in numerical properties among different indicators.

In Transformer, the attention mechanism performs attention computation on different positions of the input sequence to learn its contextual relationships and dependencies. In iTransformer, the attention mechanism is used to capture the correlation among different indicator variables, as shown in Fig. 1. The attention mechanism module performs a linear map from indicator variables to yield the query (q), key (k) and value (v) of indicator tokens, since indicator variables are normalized in their feature dimensions, the attention

mechanism computes the correlation, $\mathrm{Atten}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v})$ of q, k, and v with the following formula.

$$\mathrm{Atten}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \mathrm{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{v} \tag{8}$$

where d_k denotes the dimension of k.

III. EXPERIMENTAL EVALUATION

A. Dataset

Two datasets are adopted in this experiment. The first dataset is the real-time data of national surface water quality automatic monitoring from the China Environmental Monitoring Station [18] in Wucun, Langfang City, Hebei Province, China. It is recorded by the sensor every four hours, from Aug. 2019 to Dec. 2023. This dataset includes nine water quality indicators, *i.e.*, dissolved oxygen, total nitrogen, the potential of hydrogen, temperature, conductivity, turbidity, potassium permanganate index, ammonia, and total phosphorus. Another dataset uses the satellite remote sensing data published in the Global Satellite Precipitation Program mission [19]. It includes multi-sensor and multi-satellite information in satellite networks. Moreover, the remote sensing data is recorded every 30 minutes with a spatial resolution of 0.1° x 0.1°. The period is also from Aug. 2019 to Dec. 2023, and the variables include latitude, longitude, time, and rainfall. A typical remotely sensed rainfall image in Beijing-Tianjin-Hebei of China is shown in Fig. 3.

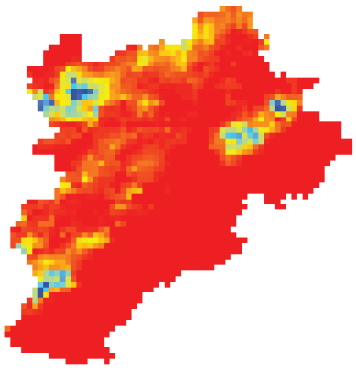

Fig. 3. A typical remotely sensed rainfall image in Beijing-Tianjin-Hebei of China

B. Evaluation Metrics

To test the prediction accuracy of CMI, mean absolute error (MAE) [20] and mean squared error (MSE) [21] are adopted. MAE and MSE are calculated as:

$$\mathrm{MAE} = \frac{1}{a}\sum_{j=1}^{a}\left|\hat{h}_j - h_j\right| \tag{9}$$

$$\mathrm{MSE} = \frac{1}{a}\sum_{i=1}^{a}\left|\hat{h}_j - h_j\right|^2 \tag{10}$$

where a denotes the number of samples. h_j and $\hat{h}_j$ denote the ground truth and predicted values of data point j.

C. Parameter Tuning

The hyperparameter selection greatly affects the prediction accuracy. CMI's hyperparameters include the length of the input sequence (S), the number of features of the encoder (D), the dimension of embedding, and the number of fusion bottleneck tokens (B).

The prediction accuracy varies significantly with S. If S is too short, the attention mechanism is not enough to capture the information, yielding lower prediction accuracy. However, if S is longer, there is too much noise or periodic information in the sequence, leading to overfitting that reduces the prediction accuracy. Table I shows the MAE and MSE values of CMI for different input sequences, and the results prove that the prediction accuracy of CMI is the optimal when S=48.

Too-small D does not capture enough information, and larger D yields a more expressive model. However, it requires more training time, computational resources, and the overfitting. During the tuning process, it is found that D has a great impact on the prediction accuracy. Table II shows MAE and MSE of CMI when $D\in$[128, 256, 512, 1024]. The results prove that CMI achieves the best prediction accuracy when D=512.

In the fusion part, the number of bottleneck tokens is the most important hyperparameter. To avoid too large computational complexity of the fusion, the number of bottleneck tokens needs to be much smaller than the number of input data tokens, Table III shows MAE and MSE of CMI with different B, and the result proves that the fusion performance is the best and the prediction result is the most accurate when B=1.

TABLE I
MAE AND MSE OF CMI WITH DIFFERENT S

S	MSE	MAE
48	**0.453**	**0.469**
72	0.465	0.471
96	0.472	0.491
120	0.494	0.517

TABLE II
MAE AND MSE OF CMI WITH DIFFERENT D

D	MSE	MAE
128	0.464	0.479
256	0.453	0.469
512	**0.438**	**0.463**
1024	0.467	0.48

TABLE III
MAE AND MSE OF CMI FOR DIFFERENT B

B	MSE	MAE
1	**0.412**	**0.442**
2	0.433	0.459
3	0.428	0.454
4	0.425	0.452
5	0.452	0.465

TABLE IV
COMPARISON OF PREDICTION RESULTS OF CMI, LMF-ITRANSFORMER, TFN-ITRANSFORMER, AND ITRANSFORMER

Prediction Length	CMI		LMF-iTransformer		TFN-iTransformer		iTransformer	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
96	**0.407**	**0.425**	0.421	0.448	0.423	0.450	0.413	0.436
128	**0.453**	**0.469**	0.475	0.491	0.476	0.490	0.478	0.480
160	**0.513**	**0.517**	0.518	0.517	0.529	0.523	0.523	0.517
192	**0.556**	**0.542**	0.557	0.542	0.584	0.559	0.579	0.556
228	**0.613**	**0.576**	0.636	0.598	0.657	0.603	0.654	0.598
256	**0.652**	**0.602**	0.674	0.612	0.705	0.629	0.715	0.628
288	**0.693**	**0.624**	0.729	0.644	0.773	0.663	0.767	0.655
320	**0.734**	**0.637**	0.745	0.650	0.788	0.673	0.831	0.687
352	**0.774**	**0.664**	0.808	0.681	0.827	0.696	0.887	0.714
384	**0.840**	**0.698**	0.877	0.716	0.896	0.723	0.950	0.746
425	**0.881**	**0.717**	0.892	0.722	0.900	0.722	1.019	0.777
450	**0.886**	**0.715**	0.915	0.736	0.925	0.728	1.062	0.796
475	**0.925**	**0.732**	0.937	0.741	0.976	0.760	1.109	0.814
500	**0.906**	**0.706**	0.975	0.765	0.936	0.715	1.131	0.818
512	**0.857**	**0.703**	0.917	0.736	0.952	0.740	1.148	0.823

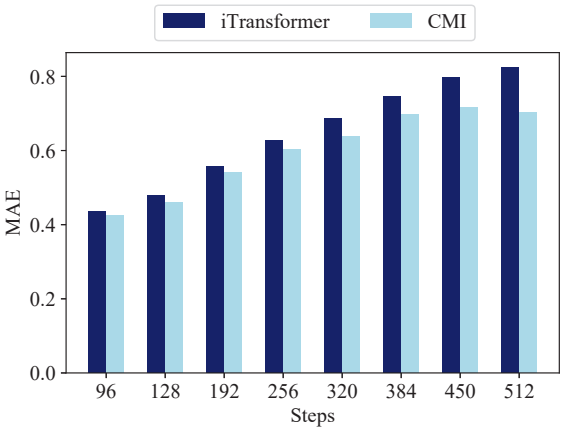

Fig. 4. MAE of CMI and iTransformer

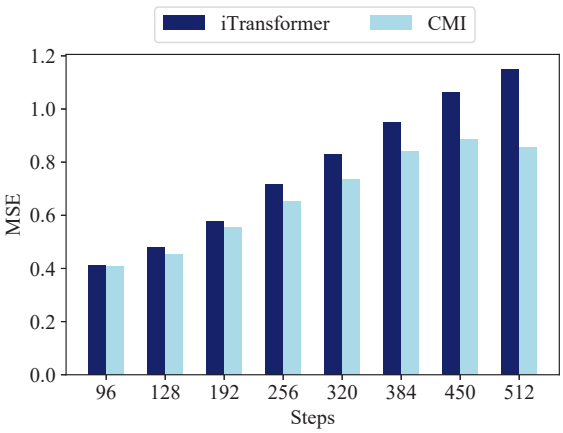

Fig. 5. MSE of CMI and iTransformer

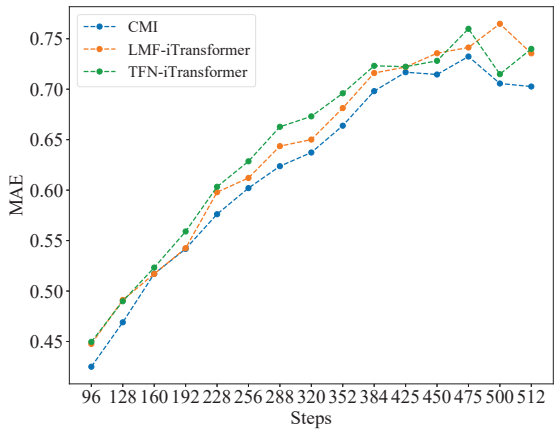

Fig. 6. MAE of CMI, LMF-iTransformer, and TFN-iTransformer

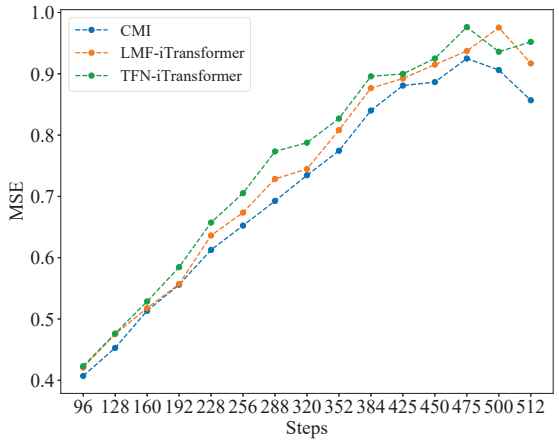

Fig. 7. MSE of CMI, LMF-iTransformer, and TFN-iTransformer

D. Comparison of Experimental Results

To demonstrate the fusion effect of water quality time series and remotely sensed rainfall data on the accuracy of water quality prediction, we first compare CMI with iTransformer. The comparison of MAE and MSE is shown in Figs. 4 and 5. The iTransformer model only adopts the water quality time series data, and CMI adopts both the water quality time series, and spatially and temporally aligned remotely sensed rainfall images. In addition, we compare CMI with two commonly used fusion models, including Low-rank Multimodal Fusion (LMF) [22] and Tensor Fusion

Network (TFN) [23]. The comparison of MAE and MSE of different models is shown in Figs. 6 and 7. Table IV shows MAE and MSE of each model when the prediction steps are in a set of (96, 128, . . . , 512), respectively. The results show that compared with LMF and TFN, CMI yields the highest prediction accuracy. In addition, CMI's prediction accuracy of water quality time series fused with remotely sensed rainfall images is on average 13% higher than that with only the time series.

IV. CONCLUSIONS

Water quality is affected by meteorological factors in addition to the water environment itself. Existing water quality prediction methods only take water quality historical indicator data as the input. However, there are many other factors that affect water quality indicators, such as meteorology and pollutants. Therefore, considering only historical data on water quality is not sufficient for accurate prediction, fusion of data from different modalities is needed. This work proposes a novel multi-indicator water quality prediction model called CMI, which combines the ConvNeXt V2, Multimodal bottleneck transformer (MBT) and Itransformer. ConvNeXt V2 is integrated to learn features of remotely sensed rainfall images and align them with the feature dimensions of time series. MBT is used to learn influences of the time series and rainfall images, and fuse their respective features. Finally, the fused new features are fed into the iTransformer for prediction. Experimental results with real-life water quality time series and remotely sensed rainfall images prove that CMI outperforms other state-of-the-art fusion algorithms including low-rank lultimodal fusion and tensor fusion network. CMI's accuracy of water quality prediction by fusing time series and rainfall images is 13% higher on average than that with only water quality time series.

In the future, we intend to further explore the impact of pollutants on water quality and integrate pollution indicators for more accurate prediction.

REFERENCES

[1] J. Bi, H. Yuan, S. Li, K. Zhang, J. Zhang, and M. Zhou, "ARIMA-Based and Multiapplication Workload Prediction With Wavelet Decomposition and Savitzky–Golay Filter in Clouds," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 4, pp. 2495-2506, Apr. 2024.

[2] H. Yuan, S. Wang, J. Bi, J. Zhang, and M. Zhou, "Hybrid and Spatiotemporal Detection of Cyberattack Network Traffic in Cloud Data Centers," *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2024.3360294.

[3] Z. Ma, H. Zhang, and J. Liu, "MM-RNN: A Multimodal RNN for Precipitation Nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-14, Apr. 2023.

[4] J. Geng, C. Yang, Y. Li, L. Lan, and Q. Luo, "MPA-RNN: A Novel Attention-Based Recurrent Neural Networks for Total Nitrogen Prediction," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6516-6525, Oct. 2022.

[5] R. Jin, Z. Chen, K. Wu, M. Wu, X. Li, and R. Yan, "Bi-LSTM Based Two-Stream Network for Machine Remaining Useful Life Prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-10, Apr. 2022.

[6] J. Bi, Z. Guan, H. Yuan, and J. Zhang, "Improved Network Intrusion Classification with Attention-assisted Bidirectional LSTM and Optimized Sparse Contractive Autoencoders", *Expert Systems with Applications*, vol. 244, pp. 1–13, Jun. 2024.

[7] C. Ma, Y. Zhao, G. Dai, X. Xu, and S. -C. Wong, "A Novel STFSA-CNN-GRU Hybrid Model for Short-Term Traffic Speed Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3728-3737, Apr. 2023.

[8] Y. Li, L. Song, S. Zhang, L. Kraus, and T. Adcox, "TCN-Based Hybrid Forecasting Framework for Hours-Ahead Utility-Scale PV Forecasting," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 4073-4085, Sept. 2023.

[9] W. Guan, X. Song, K. Wang, H. Wen, H. Ni, Y. Wang, and X. Chang, "Egocentric Early Action Prediction via Multimodal Transformer-Based Dual Action Prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4472-4483, Sept. 2023.

[10] J. Bi, Y. Li, X. Chang, H. Yuan, and J. Qiao, "Hybrid Water Quality Prediction with Frequency Domain Conversion Enhancement and Seasonal Decomposition," *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2024, Honolulu, Oahu, HI, USA, pp. 5200-5205.

[11] J. Xu and Z. Liu, "Improving the Accuracy of MODIS Near-Infrared Water Vapor Product Under all Weather Conditions Based on Machine Learning Considering Multiple Dependence Parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, Mar. 2023.

[12] J. An, F. Yin, M. Wu, J. She, and X. Chen, "Multisource Wind Speed Fusion Method for Short-Term Wind Power Prediction," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5927-5937, Sept. 2021.

[13] N. Arsha, Y. Shan, A. Anurag, J. Aren, S. Cordelia, and S. Chen, "Attention Bottlenecks for Multimodal Fusion," *Advances in neural information processing systems*, vol. 34, pp. 14200-14213, 2021.

[14] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, and Z. Cao "mplug: Effective and Efficient Vision-language Learning by Cross-modal Skip-Connections," *arXiv preprint arXiv:2205.12005*, 2022.

[15] Y. Liu, Y. Shi, F. Mu, J. Cheng, C. Li, and X. Chen," Multimodal MRI Volumetric Data Fusion With Convolutional Neural Networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-15, Jun. 2022.

[16] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and Scaling Convnets with Masked Autoencoders" *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, Vancouver, BC, Canada, pp. 16133-16142.

[17] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. "iTransformer: Inverted Transformers are Effective for Time Series Forecasting," *arXiv preprint arXiv:2310.06625,* 2023.

[18] J. Bi, Z. Chen, H. Yuan, and Jia Zhang, "Accurate Water Quality Prediction with Attention-based Bidirectional LSTM and Encoder–decoder," *Expert Systems with Applications*, vol. 238, pp. 1–10, Mar. 2024.

[19] Z. Li, D. B. Wright, S. H. Hartke, D. B. Kirschbaum, and S. Khan, "Toward a Globally-Applicable Uncertainty Quantification Framework for Satellite Multisensor Precipitation Products Based on GPM DPR," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, Jan. 2023.

[20] H. Yuan, J. Bi, S. Li, J. Zhang, and M. Zhou, "An Improved LSTM-Based Prediction Approach for Resources and Workload in Large-scale Data Centers," *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2024.3383512, pp. 1–14, Apr. 2024.

[21] S. Cai and V. K. N. Lau, "MSE Tail Analysis for Remote State Estimation of Linear Systems over Multiantenna Random Access Channels," *IEEE Transactions on Automatic Control*, vol. 65, no. 5, pp. 2046-2061, May 2020.

[22] M. Das, D. Gupta, P. Radeva, and A. M. Bakde, "Optimized Multimodal Neurological Image Fusion Based on Low-Rank Texture Prior Decomposition and Super-Pixel Segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-9, Art no. 5010409, Apr. 2022.

[23] K. Wang, Y. Wang, X. -L. Zhao, J. C. -W. Chan, Z. Xu, and D. Meng, "Hyperspectral and Multispectral Image Fusion via Nonlocal Low-Rank Tensor Decomposition and Spectral Unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7654-7671, Nov. 2020.