Saturation Line Forecasting via a Channel and Temporal Attention-Based Network

WEI WANG\(^{1,2}\), LINCHAO LI\(^{3}\), XUWEI WANG\(^{4}\), QING LI\(^{1,2}\), SHENGYAO JIA\(^{1,2}\), AND RENYUAN TONG\(^{1,2}\)

\(^{1}\)National and Local Joint Engineering Research Center for Disaster Monitoring Technologies and Instruments, China Jiliang University, Hangzhou 310051, China
\(^{2}\)Zhejiang Provincial Key Laboratory of Intelligent Manufacturing Quality Big Data Traceability Analysis and Application, China Jiliang University, Hangzhou 310051, China
\(^{3}\)Zhejiang Pecker AI Technology Ltd., Hangzhou 310018, China
\(^{4}\)College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

Corresponding author: Qing Li (lq13306532957@163.com)

This work was supported by the Zhejiang Key Research and Development Program under Grant 2018C03040 and Grant 2021C03016.

ABSTRACT Tailings ponds are places for storing industrial waste. The saturation line is the crucial factor in quantifying the safety of tailings ponds. Existing saturation line time-series prediction methods are mainly based on statistical models or shallow machine learning models. Although these models aim to capture the time dependence of the sequence data, the channel and temporal are even unavailable in principle. To mitigate this problem, in this paper, we present a two-stage forecasting method, which embeds the channel and temporal attention into a hybrid CNN-LSTM model to predict the saturation line. The channel and temporal attention are utilized to capture subtle high-dimensional time-series dependence. In the first stage, the discrete wavelet transform (DWT) is applied to capture the refined sequence information. In the second stage, the CNN-LSTM model is utilized to learn the basic spatial and temporal features in the time series. Furthermore, the channel and temporal attention model are embedded into the CNN-LSTM model to enhance the feature-extracting ability in the channel and temporal dimensions. Consequently, our proposed model is shown to outperform classic models on multiple real-world datasets in terms of RMSE, MAPE, \(R^2\) and MAE, respectively.

INDEX TERMS Deep learning, saturation line forecasting, LSTM network, attention, wavelet transform.

I. INTRODUCTION

Tailings ponds are places to store industrial waste. The tailings pond failure is ranked top 18th in the world’s risk assessment [1]. Until now, more than 100 major tailings dam accidents have been reported that caused significant damage from 1960 to 2022 [2]. The common methods for evaluating the safety situation of tailings ponds rely on manual observation or measurement analysis utilizing different sensors, e.g., water level sensors, displacement sensors, rainfall sensors and deformation sensors. In fact, considering the variability of topography, mine construction conditions, and weather conditions, the situation of the tailings dam is complicated and changeable. There are typically tailings dams located in remote mountainous areas, their structure is intricate, and the problems associated with the breakage of the dams are almost nonlinear. This makes it impossible to directly observe the tailings pond’s stability.

When the saturation line drops by 1 meter, the safety factor of static stability increases by 0.05 or more, making it the most important factor of tailings dam stability [3]. If the saturation line is too high, the dam stability will be reduced, and even leakage, landslides, and dam failure may occur [4], [5], [6]. Hence, tailings dams’ saturation line of tailings pond is termed their Lifeline [7].

It is imperative to establish a reliable model to predict the height of the saturation line so as to evaluate the security situation of the tailings pond. However, the prediction research of tailings pond is rare. To alleviate this problem, our goal is to propose a model that can take full advantage of deep learning to fit complex data. In more detail, utilizing the hidden information of the historical saturation line value,
the value and tendency in the future can be predicted. Based on this, we propose a channel and temporal attention-based CNN-LSTM network. In our model, convolutional attention layers play important roles in extracting high-dimensional structure information and passing it on to the LSTM layers for learning time-series dependence. Furthermore, the channel-wise operation in the attention module is utilized for extracting the channel structure generated by the CNN model in high dimensions. The temporal information is crucial since it expresses the implicit time series dependencies to a large degree. Meanwhile, the channel-wise features are enhanced by the channel-wise operation. Since the situation of the tailings dam is complex, the data sequence of the saturation line is unstable and without an obvious periodic structure. The noise data contains misleading and takes up a lot of space or memory. To overcome these drawbacks, we applied the discrete wavelet transform (DWT) to decompose the saturation line into different time-frequency sequences, then remove the noise in the decomposed data according to the rigorous strategy. Through decomposition and reconstruction operation, the data is refined to show the effective time-series dependence.

In this work, taking a Chinese pond as the study area, the main contributions of our study are summarized as follows:

- Proposing a two-stage forecasting model to predict the saturation line. The first is to utilize discrete wavelet transform to denoise the data, the second stage is an effective channel and temporal attention-based CNN-LSTM network. The attention module focuses on capturing time-series dependence in a different channel of the data. Our proposed model achieves satisfying performance in terms of MAPE, RMSE, MAE and $R^2$.

- Comparing our proposed model with different hyperparameters and with other state-of-the-art models to show the superiority of our proposed model.

- Conducting the ablation studies to confirm the importance of the components of our model i.e., channel and temporal attention, DWT, LSTM.

The rest of this paper is organized as follows. Section II reviews the works on tailings pond monitoring and risk prediction methods. Section III describes the datasets for saturation line prediction. Section IV introduces the proposed approach. Section V discusses the experimental results of our proposed model. Section VII conducts the ablation study of each module. Section VIII is the discussion and conclusion of this work.

II. RELATED WORK

Recently, researchers are devoted to tailings pond monitoring [8], [9], [13], [14], [15], [16]. The researchers are mainly focusing on the stability status by monitoring data from sensors and making early warnings in time by mathematical modeling method, image recognition method, and data analysis. Huang et al. [17] conducted a tailings pond monitoring and early-warning system based on three-dimensional GIS, the response time of the safety monitoring and early warning system is less than 5 seconds. Li et al. [18] proposed GPS means to monitor the displacement of tailings dams online. Gao et al. [19] established remote sensing interpretation using high-resolution remote sensing images. Necsoiu [20] used satellite radar interferometry to monitor the tailings sedimentation. Che et al. [1] assessed the risk of tailings pond by runoff coefficient, which can simultaneously determine the safety performance of multiple tailings dams. Dong et al. [21] set up an alarm system based on the cloud platform, showing good performance in real-time monitoring. Qiu et al. [3] designed a monitoring system of saturation lines based on mixed programming.

Recently, with the advantages of handling almost any nonlinear and linear problems, regardless of low and high dimensions, neural network and machine learning methods have been effectively composed in real-time risk analysis and evaluation [22], [23], [24], [25], [26], [27], [28], [29]. However, the role of real-time monitoring cannot be equated with early warning and forecasting. In other words, risk prediction methods could help people perceive risk before it happens. With excellent ability to process time-series, classic prediction models such as Auto-Regressive Integrated Moving Average (ARIMA) and LSTM have been used in prediction problems [30], [31], [32], [33], [34]. They analyze and identify the time series information of training data and give the prediction value a few days in advance. Nevertheless, different from LSTM, the ARIMA model only gets a high score in the condition of data with linear correlation or without obvious fluctuation. With the rapid development of deep learning, CNN and LSTM have been the most popular networks. The CNN could filter out the noise data and extract important features, achieving good performance in images, speeches, and time-series [35], [36]. While the LSTM network has the ability to find the linear or nonlinear time series information from the shallow and deep network and combine it with current memory [37]. In light of this, combining LSTM with CNN may achieve better prediction performance to a large extent.

Previous prediction methods have either been shallow machine learning models or directly used LSTM. These methods only consider the temporal correlation of the data. In this paper, we first input the data after DWT refinement into the CNN. We then utilize channel and temporal-wise attention to enhance the feature extraction. Finally, we use LSTM for time series forecasting. Our scheme fully considers the temporal, spatial, and inter-channel correlation of data. Although our model is more complex, the running time is fast and can achieve real-time prediction, which benefits from the parameter-sharing mechanism of CNN.

III. DATA PREPARATION

The study site is a copper mine tailings pond in Zhejiang Province, China. The real-world historical data for this work are collected from the water level sensors from 2020-03-18 to 2021-04-29. Specifically, we collect five datasets from the
different heights of the tailings pond. The water level sensor is shown in Figure 1.

We use medium imputation techniques to fill in missing values. Due to the characteristics of the saturation line data, its value will not change abruptly. Therefore, it is reasonable that the missing value is between the data value of the previous moment and the data value of the next moment. After filling the missing value with medium value and replacing the abnormal value with medium value, 9,865 data points are collected for our prediction task. The monitoring data are continuous, which maintains a wide range of original time-series information. Notably, for the 9,865 data, 70% of the dataset is set as the training sets and 10% as the validation set. The performance of our models is confirmed on the rest 20% dataset. We do not shuffle the data as usual in traditional deep learning studies since the time-series dependence relies on the time-series order. Table 1 indicates the description of the monitoring data. The first three rows show the historical monitoring data, while the other columns show the statistical details.

In order to eliminate the impact of different data dimensions on the calculation, we apply Z-score normalization, the formula is as follows:

$$\hat{x} = \frac{x_t - \mu_t}{\sigma_t}$$

where $x_t$ is the input data, $\mu_t$ and $\sigma_t$ are the averages and standard deviation of data.

### IV. METHOD

This section first describes the overview of our model. The DWT method is indicated in Sec. IV-B, and our proposed model is explained in Sec. IV-C.

#### A. OVERVIEW

In this paper, we present a two-stage forecasting method, which embeds the channel and temporal attention into a CNN-LSTM model to predict the saturation line. In the first stage, i.e., first row in Figure 2, the DWT is applied to capture the refined sequence information. In the second stage, i.e., second row in Figure 2, the CNN-LSTM model is used to learn the spatial and temporal features in the refined time series. Furthermore, the channel and temporal attention model is utilized to enhance the feature-extracting ability.

#### B. DISCRETE WAVELET TRANSFORM

There are some classical techniques for data denoising, such as fast Fourier transform (FFT) [10], short-time Fourier transform (STFT) [11], electrical discharge machining (EDM) [12], and discrete wavelet transform (DWT) [13]. FFT and STFT cannot meet the frequency requirements of unstable state signal changes. EDM has a modal aliasing effect at a low signal-to-noise ratio.

Although the window Fourier transform (short-time Fourier transform) can partially locate the time, since the window size is fixed, it is only suitable for stationary signals with small frequency fluctuations, and not suitable for non-stationary signals with large frequency fluctuations. As a signal time-frequency analysis method, the wavelet transform (WT) can automatically adjust the window size according to the frequency. What has greatly contributed to the effectiveness of WT is the truth that it is an adaptive time-frequency analysis method which can perform multi-resolution analysis. As a result, wavelet transform is known as a microscope for analyzing and processing signals.

DWT can remove the indistinguishable noise and impurity components in the time domain, and can simultaneously examine the frequency domain and time domain characteristics. Therefore, even in the face of those non-stationary processes, DWT can transform and process well. In addition, since the mutation of our data infiltration line is generally a continuous process and does not mutate, DWT does not lose important data. In other words, the removed data is generally an outlier.

In our study, we apply the DWT to decompose the collected saturation line data of the tailings pond into 4 frequency sequences. After removing the noise in the decomposed data, the wavelets are reconstructed to obtain new integrated data for further multi-resolution study. We tested the prediction effect on multiple datasets, and DWT proved to be an effective denoising method from the results.

The WT refers to the displacement of a certain basic wavelet function by $\omega$ units, and then the inner product with the analysis signal $p(t)$ at different scales.

$$WT_{\epsilon}(\omega, \epsilon) = \frac{1}{\sqrt{n}} \int_{-\infty}^{+\infty} p(t) \phi(\frac{t-\omega}{n})dt$$

where $\epsilon$ is the scale factor (> 0) to stretch each basic wavelet $\phi(t)$. $\omega$ is the displacement. Mallat algorithm [38] provides an effective way to display DWT to process the data using
the low-high-pass filters:

\[
oL = \sum_{i=-\infty}^{\infty} T(i)\psi_l(2n-i) \tag{3}
\]

\[
oH = \sum_{i=-\infty}^{\infty} T(i)\psi_h(2n-i) \tag{4}
\]

where \( T(i) \) means the signal. \( \psi_l, \psi_h, oL, oH \) are the low-pass filter, high-pass filter, output of low-pass filter, and output of high-pass filter, respectively. Notably, in the wavelet domain, the coefficient corresponding to the effective signal is large, and the coefficient corresponding to the noise is small. As a result, the noise can be removed by the threshold. In this paper, we apply the common \textit{rigrsure} threshold in DWT:

\[
g(k) = [\text{sort}|t|]^2, \quad (k = 0, 1, \ldots, N - 1) \tag{5}
\]

In the equation, the absolute value of each signal is achieved and then sorted, and the square of each number is taken to obtain a new signal sequence.

\[
\gamma_t = \sqrt{g(t)}, \quad (t = 0, 1, \ldots, N - 1) \tag{6}
\]

\[
\text{Risk}(t) = \frac{(N - 2t + \sum_{j=1}^{t} g(j) + (N - t)f(N - t))}{N} \tag{7}
\]

\[
\gamma_t = \sqrt{\text{Risk}^{\text{min}}} \tag{8}
\]

The \( t \) is the signal, \( \gamma_t \) is the threshold and \( \text{Risk}(t) \) is the generated risk. Take the minimum \( g(t) \) corresponding to all risks \( r(K) \) to get the final threshold \( \gamma_t \).

The 3-level decomposition and the reconstruction process of DWT with Mallat algorithm are shown in Figure 4(a) and Figure 4(b), respectively. From Figure 4(a), we can see that after decomposing the signal into three different levels. In more detail, at the first level, the original signal \( T \) is decomposed to the detail coefficients \( oL_1 \) and \( oH_1 \). Then the achieved \( oL_1 \) is decomposed to the other two coefficients \( oH_1 \) and \( oL_1 \) at the second level. The decomposition process does not end until the set number of \( n \)-level steps is reached. Figure 4(b) illustrates the process of de-noise and reconstruction. The noises are shown with small wavelet coefficients, while the useful signals are shown with small wavelet coefficients. The time-series signal \( T \) passes through the low-pass filter \( oL_1 \) and high-pass filter \( oH_1 \) to removing the wavelet coefficients of lower amplitude and restore the wavelet coefficients of higher amplitude to achieve the effect of noise reduction. Subsequently, the wavelet reconstruction and integration process is applied to all of these coefficients. Employing the coefficient \( oL_3 \), the low frequency and high amplitude \( rL_3 \) is reconstructed. As shown in \( rL_3 \) in Figure 4, the sequences become smooth, showing the refined sequence patterns.

C. STRUCTURE OF THE PROPOSED MODEL

Our study aims to develop the construction of a prediction system for forecasting the saturation line utilizing state-of-the-art LSTM and CNN networks. What has been devoted to the popularity of the convolutional layer is the fact that it is good at extracting and recognizing as well as identifying the structures of the time series in the monitoring data, while the LSTM networks achieve good performance in detecting long-short-term dependence. In light of this, the principle idea of our study is to combine the advantages of CNN and LSTM.

We first build a baseline model, namely CNN-LSTM, which utilizes one CNN and one LSTM model. In a further study, our proposed model improves the baseline with one more LSTM model together with two channel and temporal attention models. As we mentioned before, our proposed model is a two-stage model, where the first part is the DWT process, and the second part is the time-series prediction model. The convolutional layers encode the time-series information as a high-dimensional structure, while the LSTM layer decodes the information from convolutional layers for...
time-series dependence. The channel and temporal-wise features are enhanced by the attention model. Our proposed model is shown in Figure 2.

Specifically, our proposed model includes one convolutional layer filters of 32, a max-pooling layer filters of 2, two LSTM layers of 25 and 50, a flatten layer, and a fully-connected layer in order. Different parameters of CNN and LSTM are compared for further study in Table. 4. Furthermore, the attention module is embedded in the baseline CNN-LSTM model. The details of the embedded channel and temporal attention are described in Sec. IV-D.

**D. CHANNEL AND TEMPORAL ATTENTION MODULE**

In this section, we explain the channel and temporal attention we utilized in this paper. The channel-wise operation is utilized for extracting the channel structure generated by the CNN model in high dimensions.

For the channel-wise features, the spatial information of a feature map is aggregated by the average-pooling and max-pooling operations, generating two different spatial descriptors: $F_{c_{avg}}$ and $F_{c_{max}}$, which represent average-pooled features and max-pooled features respectively.

$$F^c = \sigma(FC(AvgPool(F) + FC(Maxpool(F))))$$
$$= \sigma(W_1(F^c_{avg}) + W_1(F^c_{max})), \quad (9)$$

where $\sigma$ indicates the sigmoid function and $FC$ indicates the shared fully connected layer with $Relu$ function.

The final channel-wise attention is formulated as:

$$F' = F^c \otimes F,$$ \quad (10)

where $\otimes$ means element-wise product.

For the time-series prediction task, the temporal information is crucial since it expresses the implicit time series dependencies to a large degree. For the temporal-wise features, the channel information of a feature map by using two pooling operations is generated as two maps.

$$F^t = \sigma(Conv_{7 \times 7}(Concat[Avgpool(F), Maxpool(F)]))$$
$$= \sigma(Conv_{7 \times 7}(Concat[F^t_{avg}, F^t_{max}]), \quad (11)$$

where $Conv_{7 \times 7}$ represents the convolution kernel with the size of $7 \times 7$.

The overall channel and temporal-wise attention is formulated as:

$$F'' = F' \otimes F'.$$ \quad (12)

To summarize the whole method, we describe the proposed channel and temporal attention-based network in Algorithm 1. The input of the algorithm includes raw data $D$, fully connected layer $FC$, $7 \times 7$ convolution layer, and discrete wavelet transform $DWT$. The output of the algorithm is the forecast result.

**V. EXPERIMENT AND RESULTS**

**A. METRICS**

The prediction performance of our proposed model is evaluated by the root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute percentage (MAE), and coefficient of determination ($R^2$). In fact, RMSE meets
Algorithm 1 Channel and Temporal Attention-Based Network

**Input:** Raw data $D$, fully connected layer $FC$, $7 \times 7$ convolution layer, and discrete wavelet transform $DWT$

**Output:** Forecast result $R$

1. Normalize raw data by Eq.(1)
2. Denoise raw data: $D_n = DWT(D)$
3. Extract features by convolutional layer: $F = CNN(D_n)$
4. Compute channel-wise attention $F_c$ by Eq.(10)
5. Compute temporal-wise attention $F_t$ by Eq.(11)
6. Compute channel and temporal-wise attention $F_a$ by Eq.(12)
7. Compute time series dependence: $R = LSTM(F_a)$
8. return $R$

where $y_t$ represents the true value, $\hat{y}_p$ represents predicted saturation line value, $\bar{y}$ represents average of true value, and $n$ is the count of data.

**VI. EXPERIMENT**

Our proposed model is evaluated and compared to other models to show the prediction performance. The prediction

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_t - \hat{y}_p)^2}$$  \hspace{1cm} (13)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_t - \hat{y}_p|}{y_t} \times 100\%$$  \hspace{1cm} (14)$$

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_t - \hat{y}_p)^2}{\sum_{i=0}^{n}(y_t - \bar{y})^2}$$  \hspace{1cm} (15)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_t - \hat{y}_p|$$  \hspace{1cm} (16)$$
performance of our proposed model is shown in Table 2, where NO.8, NO.13, NO.17, NO.21, NO.28, NO.33 mean the different stations of the saturation line mentioned above. The deep and abstract features of the convolutional layer learned may be different from the ordinary time-series information from the raw data. This is obviously a disadvantage when the monitoring data contains only simple or even linear information. While using one convolutional layer and two LSTM layers together with the channel and temporal attention can capture the long-short-term data dependencies to a significant degree from the result.

| Metrics  | NO.8  | NO.13 | NO.17 | NO.21 | NO.28 | NO.33 |
|----------|-------|-------|-------|-------|-------|-------|
| RMSE     | 0.0209| 0.0300| 0.0336| 0.0170| 0.0123| 0.0366|
| MAPE     | 3.346 | 3.316 | 3.207 | 1.589 | 3.221 | 3.432 |
| R²       | 0.969 | 0.974 | 0.937 | 0.981 | 0.951 | 0.892 |
TABLE 3. Performance comparison of different machine learning and deep learning models. For a fair comparison, we only compare the runtime of deep models. **ATR** means the average training time for each epoch, and **ATE** means the average testing time for each epoch.

| Type       | RMSE  | MAPE  | $R^2$ | MAE   | Runtime (s) | ATR (ms) | ATE (ms) | Model      |
|------------|-------|-------|-------|-------|-------------|----------|----------|------------|
| Shallow    | 0.132 | 4.542 | 0.548 | 0.101 | —           | —        | —        | SVR        |
|            | 0.141 | 4.312 | 0.489 | 0.073 | —           | —        | —        | DTR        |
|            | 0.251 | 4.186 | 0.839 | 0.193 | —           | —        | —        | RFR        |
| Deep model | 0.0504| 3.744 | 0.798 | 0.0552| 44.08       | 220.4    | 146.5    | MLP        |
|            | 0.0461| 3.734 | 0.769 | 0.0571| 38.32       | 197.2    | 137.95   | DBN        |
|            | 0.0308| 3.645 | 0.864 | 0.0402| 47.54       | 237.7    | 152.5    | RNN        |
|            | 0.0221| 3.602 | 0.879 | 0.0366| 63.35       | 317.7    | 212.6    | GRU        |
|            | 0.0214| 3.596 | 0.887 | 0.0338| 77.08       | 385.4    | 289.0    | LSTM       |
|            | 0.0220| 3.562 | 0.953 | 0.0210| 45.21       | 226.1    | 129.8    | Dual-stage attention |
|            | 0.0269| 3.982 | 0.931 | 0.0239| 26.18       | 130.9    | 86.5     | PSTA-TCN   |
|            | 0.0216| 3.391 | 0.962 | 0.0156| 58.06       | 290.3    | 202.2    | Graph attention LSTM |
|            | 0.0209| 3.346 | 0.969 | 0.0133| 25.49       | 127.4    | 82.9     | Ours       |

TABLE 4. Prediction cases using different hyperparameters in our proposed model.

| Batch Size | Convolution kernel Size | Pooling Size | LSTM Units | RMSE  | MAPE  | $R^2$ | MAE   | Runtime (s) |
|------------|-------------------------|--------------|------------|-------|-------|-------|-------|-------------|
| Case 1     | 16                      | 16           | 2          | [50,50] | 0.0296| 3.322 | 0.972  | 0.0135 | 51.32       |
| Case 2     | 16                      | 32           | 2          | [25,75] | 0.0385| 3.513 | 0.902  | 0.0315 | 50.52       |
| Case 3     | 32                      | 16           | 2          | [25,50] | 0.0411| 3.352 | 0.899  | 0.0280 | 50.87       |
| Case 4     | 32                      | 16           | 4          | [25,50] | 0.0514| 4.114 | 0.792  | 0.0339 | 21.13       |
| Case 5     | 64                      | 32           | 2          | [25,50] | 0.0209| 3.346 | 0.969  | 0.0133 | 25.49       |
| Case 6     | 64                      | 32           | 2          | [50,75] | 0.0208| 3.324 | 0.971  | 0.0145 | 49.37       |
| Case 7     | 64                      | 32           | 4          | [50,75] | 0.0504| 4.011 | 0.801  | 0.0309 | 21.88       |
| Case 8     | 128                     | 32           | 4          | [25,50] | 0.0521| 4.281 | 0.784  | 0.0367 | 25.12       |
| Case 9     | 128                     | 32           | 2          | [25,50] | 0.0311| 3.501 | 0.932  | 0.0223 | 37.49       |

The scatter plots of raw data and predicted saturation line is illustrated in Figure 10, which helps show the prediction performance intuitively.

To show the superiority of our proposed model, we apply comparative studies with other state-of-the-art machine learning, deep learning, and attention models, including the support vector regression (SVR), decision tree regression (DTR), random forest regression (RFR), deep belief network (DBN), multilayer perception (MLP), single GRU, simpleRNN, LSTM, Dual-stage attention [39], PSTA-TCN [40] and Graph attention LSTM [41] models on NO.8 dataset. Table 3 presents the RMSE, MAPE, MAE, and $R^2$ score of these models in our experiments, which demonstrates that our proposed model significantly outperforms the others in $R^2$. Besides, the runtime for the whole training process, the average training time for each epoch (ATR), and the average testing time for each epoch (ATE) are less than other deep learning models.

In order to build the complete saturation line prediction model and show the reliability of our proposed model together with the parameters set, we compared different hyperparameters such as batch size, filters in the convolutional layers, max-pooling size, and number of LSTM cells in our experiments. Table 4 lists the different situations of combing multiple hyperparameters. Considering several evaluation metrics and running time, we choose the design shown in Case 5. Specifically, in terms of the evaluation metrics used in this task, although Case 1 and Case 6 achieve slightly better performance than that in Case 9, the Runtime is almost twice. Predictions will perform worse in real-time when the running time is excessive, especially when there is a large amount of data. The disadvantage is more pronounced for a large amount of data, and this incurs no loss of generality. Case 4 needs the least Runtime but achieves low accuracy. In addition, we use MAE to do 10-fold cross-validation for these 9 sets of parameters, as shown in Figure 8. And we calculated the average MAE of cross-validation for each set of parameters as shown in Table 4. The results demonstrate that case 5 achieves the best performance in MAE cross-validation. To be clear, according to Case 5, the implementation is: the batch size is
The prediction scatters of saturation line at different positions utilizing the baseline model (CNN-LSTM).

VII. ABLATION STUDY

We conduct the ablation study to evaluate the effectiveness of DWT and channel and temporal-wise attention. Since our proposed model includes the DWT process, one convolutional layer, two LSTM layers, and two channel and temporal attention modules, we evaluate these experiments by removing these modules. The results display in Table 5. Furthermore, we compare the fitting results with the baseline model (one CNN and one LSTM). It can be seen that our proposed model (Figure 10) achieves more effective predictions than that utilizing the baseline model (Figure 9), especially in predictions NO.17, NO.21, and NO.28. The results can be also confirmed by comparing the prediction curve using our proposed model and baseline model shown in Figure 6 and Figure 7. This is because although the long-term and short-term dependence and hidden time series information can be discovered from the data, the prediction accuracy is greatly affected due to the presence of noise in the data. Furthermore, the channel and temporal-wise features are ignored by the simple CNN-LSTM model.

To overcome the drawbacks of the baseline model that cannot de-noise the raw data, we applied the DWT to decompose the saturation line into different time-frequency sequences and remove the random noise. Subsequently, the data after noise reduction is trained by our proposed model. With the help of enhancing the channel and temporal features, the results of all positions are shown in Table 5. This once again proves that the DWT method can remove a large amount of useless information, thereby assisting our model to more accurately explore the time series information hidden between the data. It also can be illustrated from Table 5 and the comparison of Figure 9 and Figure 10 that our proposed model achieves better performance than the baseline model at all saturation line stations.

For example, on dataset No.8, the baseline achieves the $R^2$ of 0.931, while our proposed model has the value of 0.969. When removing the DWT, the model has a value of 0.942 in terms of $R^2$; when removing the attention module, the model has a value of 0.953 in terms of $R^2$; when removing the LSTM, the $R^2$ drops from 0.969 to 0.961. The results show that all the components are crucial for our proposed model. In addition, we can see from Table 5 that when DWT was removed from the model, $R^2$ decreases the most. Multiple
experiments show that DWT plays the most critical role of all the modules.

**VIII. DISCUSSION AND CONCLUSION**

In this work, we applied a new method to predict the safety of the tailings pond according to the saturation line using our proposed model, which is also first used in tailings pond risk prediction. Compared with the traditional methods, the risk evaluation method of tailings ponds has the characteristics of high accuracy and high real-time performance.

The contributions of this work are three fold: First, proposing an effective channel and temporal attention-based CNN-LSTM network to predict the saturation line, which achieves satisfying performance in terms of MAPE, RMSE, MAE, and $R^2$. Second, comparing our proposed model with different hyperparameters and with other state-of-the-art
models. Third, conducting ablation studies to confirm the components of our model. The wavelet transform method is applied to overcome the shortcomings that the original CNN-LSTM model could not de-noise the data. The wavelet transform decomposes the data into layers of wavelets, selects the recurrence threshold to de-noise the decomposed wavelet, and then reconstructs them, subsequently feeding the reconstructed useful signals to our attention-based model to obtain better prediction results.

In the tailings pond risk prediction task, these experiments consequently provide applicability of the proposed model. Additionally, our model can also be used to predict water levels, weather, and air quality as time-series predictions. The model is evidently capable of not only extracting and recognizing spatial and time series structures but also identifying long-term and short-term series information.

Our method applies one factor, and in the future, we will focus on more factors of the safety monitoring parameters of the tailings pond, such as the underground displacement, ground displacement, and dry beach length. We should also build a risk level that corresponds to the tailings pond monitoring so that our future work more intuitively reflects the safety of the tailings pond.

REFERENCES

[1] D. Che, A. Liang, X. Li, and B. Ma, “Remote sensing assessment of safety risk of iron tailings pond based on runoff coefficient,” Sensors, vol. 18, no. 12, p. 4373, Dec. 2018.
[2] World Information Service on Energy Uranium Project. Chronology of Major Tailings Dam Failures. Accessed: May 19, 2022. [Online]. Available: https://www.wise-uranium.org/mdaf.html
[3] P. Qiu, X. Yuan, S. Gan, J. Zhang, and M. Yang, “Monitoring system of saturation line based on mixed programming,” in Proc. 2nd Int. Conf. Comput. Eng., Inf. Sci. Appl. Technol. (ICCIAT), 2017, pp. 146–149.

A. Medyńska-Juraszek and C. Kabala, “Heavy metal pollution of forest soils affected by the copper industry,” J. Ecol., vol. 17, no. 3, pp. 1–11, Jul. 2012.

A. Medyńska-Juraszek and L. Kuchar, “Carbon sequestration rates in organic layers of soils under the grey poplar (Populus canescens) stands impacted by heavy metal pollution,” in Functions of Natural Organic Matter in Changing Environment. Dordrecht, The Netherlands: Springer, 2013, pp. 365–369.

S. Fiaschi, M. Mantovani, S. Fiaschi, A. Pasuto, and M. Floris, “Testing the potential of Sentinel-1A TOPS interferometry for the detection and monitoring of landslides at local scale (Vetino, regioni, Italy),” Environ. Earth Sci., vol. 76, no. 14, p. 492, Jul. 2017.

F. Wang, Y. A. N. K. Teng, and X. U. Zhi-Sheng, “Stability analysis of tailings dam based on saturation line matrix,” Rock Soil Mech., vol. 30, no. 3, pp. 840–844, 2009.

Z. Qi and L. Xuelong, “Big data epoch: Challenges and opportunities for geology,” Bull. Chin. Acad. Sci., vol. 33, no. 8, pp. 825–831, 2018.

H. E. Chaoyang, J. U. Nengpan, and H. Jian, “Automatic integration and analysis of multi-source monitoring data for geo-hazard warning,” J. Eng. Geol., vol. 22, no. 3, pp. 405–411, 2018.

M. Frigo and S. G. Johnson, “FFTW: An adaptive software architecture for the FFT,” in Proc. Int. Conf. Acoust., Speech, Signal Process., vol. 3, May 1998, pp. 1381–1384.

M. K. Kyiymik, I. Güler, A. Dizibüyük, and M. Akin, “Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application,” Comput. Biol. Med., vol. 35, no. 7, pp. 603–616, Oct. 2005.

K. H. Ho and S. T. Newman, “State of the art electrical discharge machining (EDM),” Int. J. Mach. Tools Manuf., vol. 43, no. 13, pp. 1287–1300, Oct. 2003.
[39] Y. Xiao, H. Yin, Y. Zhang, H. Qi, Y. Zhang, and Z. Liu, “A dual-stage attention-based Conv-LSTM network for spatio–temporal correlation and multivariate time series prediction,” Int. J. Intell. Syst., vol. 36, no. 5, pp. 2036–2057, May 2021.

[40] J. Fan, K. Zhang, Y. Huang, Y. Zhu, and B. Chen, “Parallel spatio–temporal attention-based TCN for multivariate time series prediction,” Neural Comput. Appl., vol. 2021, pp. 1–10, May 2021.

[41] T. Zhang and G. Guo, “Graph attention LSTM: A spatiotemporal approach for traffic flow forecasting,” IEEE Intell. Transp. Syst. Mag., vol. 14, no. 2, pp. 190–196, Mar. 2022.

WEI WANG received the M.S. degree from the School of Mechanical and Electrical Engineering, China Jiliang University. His research interests include machine learning, image processing, and software development.

LINCHAO LI received the M.S. degree from China Jiliang University. He is currently a Machine Learning Algorithm Engineer. His research interests include computer vision and image processing.

XUWEI WANG received the M.S. degree from the College of Civil Engineering and Architecture, Zhejiang University, where he is currently pursuing the Ph.D. degree in computer science and technology. He has five years working experience in computer vision field and obtains three authorized invention patents. His research interests include AI, deep learning, and computer vision.

QING LI was the Dean of the School of Mechanical and Electrical Engineering, China Jiliang University, where he is currently a Professor. He is also the Executive Director of China’s Metrological Testing Association. He was selected as a Famous Teacher of the National “Ten-Thousand Talents Program,” in 2017. His current research interests include dynamic measurement and control and sensing technology. He has presided over the completion of more than 60 science and technology projects.

SHENGYAO JIA received the B.E. degree in measurement and control technology and instrumentation and the M.S. degree in detection technology and automation from China Jiliang University, Hangzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree in agricultural electrification and automation from Zhejiang University, Hangzhou, in 2015. He is currently a Lecturer with China Jiliang University. His current research interests include energy harvesting systems, sensors and measuring technology, and embedded systems.

RENYUAN TONG received the Graduate degree from East China Normal University. He is currently an Associate Professor with the College of Mechanical and Electrical Engineering, China Jiliang University. His research interest includes detection technology.