Transportation Geotechnics 40 (2023) 100957

Contents lists available at ScienceDirect

Transportation Geotechnics

journal homepage: www.elsevier.com/locate/trgeo

Data-driven analysis on the subbase strain prediction: A deep data augmentation-based study

Hui Yao[a], Shibo Zhao[a], Zhiwei Gao[b], Zhongjun Xue[c], Bo Song[c], Feng Li[d], Ji Li[e], Yue Liu[f], Yue Hou[e,a,*], Linbing Wang[g,*]

[a] Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, No.100 Pingleyuan, Chaoyang District, Beijing, China
[b] James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK
[c] Beijing Key Laboratory of Road Materials and Testing Technology, Beijing Road Engineering Quality Supervision Station, Beijing, China
[d] School of Transportation Science and Engineering, Beihang University, No.9 Nansan Street, Changping District, Beijing 102206, China
[e] Department of Civil Engineering, Faculty of Science and Engineering, Swansea University, UK
[f] Research Institute of Urbanization and Urban Safety, School of Civil and Resource Engineering, University of Science and Technology Beijing, 30 Xueyuan Road, Beijing 100083, China
[g] School of Environmental, Civil, Mechanical and Agricultural Engineering, University of Georgia, Athens, GA, USA

ARTICLE INFO

Keywords:
Subbase strain development
Intelligent analysis
Data augmentation
Model interpretability
Deep analysis

ABSTRACT

The service quality of the subbase may affect the overall road performance during its service life. Thus, monitoring and prediction of subbase strain development are of great importance for civil engineers. In this paper, a method based on the time-series augmentation was employed to predict the subbase strain development. The time-series generative adversarial network (TimeGAN) model was implemented to perform the augmentation of time-series data based on the original monitored data. The augmented data was trained through deep learning network to learn the feature correlation of the subbase strain. The effectiveness of TimeGAN on the prediction accuracy was evaluated through the Attention-Sequence to Sequence (Attention-Seq2seq) model, and temporal convolution network-adaptively parametric rectifier linear units (TCN-APReLU) model. Results indicated that the TimeGAN network could capture sufficient information from the time-series monitored data of subbase strain development so that the corresponding augmented data matches well with the original data, which improves the prediction accuracy. It is also discovered that the combination of TimeGAN and TCN-APReLU appropriately predict the subbase strain development based on the original monitored data.

Introduction

The resilience and deformation of the subbase are the key issues for the service life [1]. Thus, accurate and timely predictions of the strain development in the subbase are crucial for health monitoring and structural rehabilitations [2–5]. With the fast development of computer and sensor technology, long-term monitoring data can be obtained from embedding sensor devices, which provides the possibility for civil engineers to investigate the development of subbase based on the monitored temporal data.

Deep learning-based neural networks can be used as powerful tools to learn the data characteristics based on large amounts of monitored data. Bansal et al. (2022) developed machine learning models to monitor and predict ternary blended concrete strength using the measured data from piezo sensors [6]. Tabrizi et al. (2021) aimed to develop a reliable and accurate pavement surface temperature prediction tool using machine learning techniques [7]. Advanced machine learning techniques provide greater efficiency and lower computational effort compared to the traditional mechanistic theory of numerical analysis [8]. As the machine learning approaches normally learn the data features through iterative computations, the model structure and the amounts of monitored data are the key points for the quality of analysis [9].

Civil engineers sometimes face the problems of limited monitored data due to the limitations of monitor instruments, external environmental conditions, etc. which may result in a low computation accuracy

[*] Corresponding authors at: Department of Civil Engineering, Faculty of Science and Engineering, Swansea University, UK (Yue Hou); School of Environmental, Civil, Mechanical and Agricultural Engineering, University of Georgia, USA (Linbing Wang).

E-mail addresses: yue.hou@swansea.ac.uk (Y. Hou), Linbing.Wang@uga.edu (L. Wang).

https://doi.org/10.1016/j.trgeo.2023.100957
Received 17 October 2022; Received in revised form 27 December 2022; Accepted 8 February 2023
Available online 11 February 2023

in the corresponding machine learning-based approach. To solve this problem, data augmentation approaches can be used to increase the size of training data [10]. For instance, transforming and rotation were used as the data augmentation method to create new data in the field of computer vision [11,12]. Recently, data augmentation techniques were gradually applied in time series data analysis. Wu et al. (2021) designed a hybrid time series data augmentation framework to solve the problem of class-imbalance and insufficient sample size in the application of deep learning model [13]. Jeong et al. (2021) proposed a novel time-warping data augmentation method reflecting the characteristics of the sensor signal and can preserve the label of the augmented signal by generating partially occluded data of the accelerometer signals [14]. Almonacid et al. (2013) used an artificial neural network (ANN) for the ambient temperature hourly time series data [15]. Taylor et al. (2022) developed a new neural network based on a multi-head architecture and data augmentation [16]. Fu et al. (2019) proposed a conditional generative adversarial net (CGAN) to learn and simulate time series data [17]. In previous research, the authors have evaluated the prediction performance on the long-term subbase strain data of three deep learning methods, including the Long-Short Term Memory neural network (LSTM) model, Bidirectional LSTM-Convolution Neural Network (BiLSTM-CNN) model and Temporal Convolution Network (TCN) model [18]. Results show that the deep learning methods exhibited a good performance for the long-term subbase strain data analysis [18]. However, there are still several problems: firstly, the pre-processing of time-series data may not be sufficient due to the lack of sufficiently large samples; secondly, the monitored sparse data may lead to worse performances of models.

To solve these problems, in this study, data-based approaches, including the random forest algorithm and smoothing operation were conducted on the original monitoring data. A widely used TimeGAN model for data augmentation was employed to augment the original monitored data of subbase strain. Then, the effectiveness of the data augmentation techniques was evaluated using two established deep learning models by other researchers, i.e., TCN-APReLU and Seq2seq. Last, the augmentation results of two other data augmentation methods, variational autoencoder (VAE) and Wasserstein Generative Adversarial Network (GAN) with Gradient Penalty (WGAN-GP), were analyzed and compared with that of the TimeGAN. The combination of TimeGAN and TCN-APReLU applied in this paper can be used for data augmentation of infrastructure monitoring data and the prediction of subbase strain. The whole technical route of this study is shown in Fig. 1.

Methodologies

Data augmentation method-TimeGAN

Data augmentation plays an important role in enlarging the data quantity and improving prediction accuracy. Traditional data augment methods focused on the relationship between the time-independent features. For classic GAN, a random noise vector with a gaussian distribution was generated and added to the train data in the generator network. The discriminator was trained to maximize the probability of assigning the correct label to samples from the generator and the train data [19–21]. The train function was shown in equation (1) [19].

$$\underset{G}{min}\underset{D}{max}V(D,G)=\mathbb{E}_{x\ p(x)}[logD(x)]+\mathbb{E}_{z\ p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

where, $p(x)$ is the training data distribution, $p_z(z)$ is the prior distribution of the generator network, and z is a noise vector sampled from the model distribution $p_z(z)$.

The temporal feature of the time-series data poses a serious challenge to the traditional data augment method. The data augmentation of a time-series data should not only deal with the feature within a time point, but also capture the complex property of those feature across the time. The TimeGAN model, which consists of four different network components, was employed for the time-series data augmentation [22]. The components of TimeGAN can be classified as the autoencoding

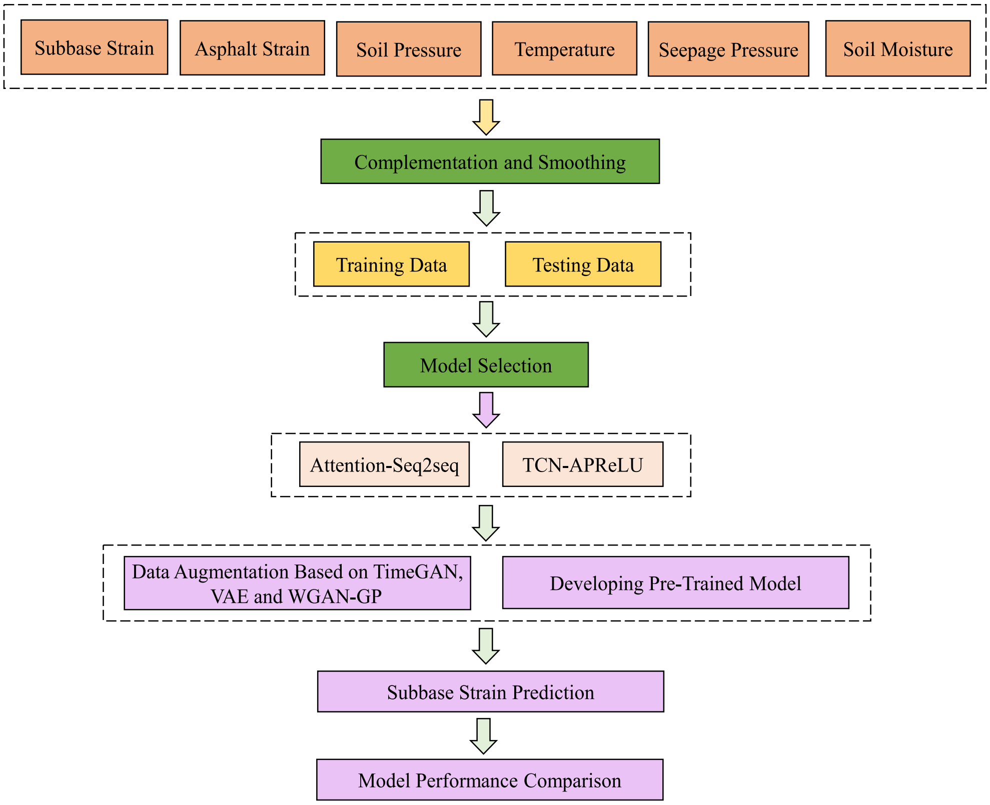

Subbase Strain
Asphalt Strain
Soil Pressure
Temperature
Seepage Pressure
Soil Moisture
Complementation and Smoothing
Training Data
Testing Data
Model Selection
Attention-Seq2seq
TCN-APReLU
Data Augmentation Based on TimeGAN, VAE and WGAN-GP
Developing Pre-Trained Model
Subbase Strain Prediction
Model Performance Comparison

Fig. 1. Flowchart of this study.

group (embedding function and recovery function) and the adversarial group (sequence generator and sequence discriminator). The structure of the TimeGAN is shown in Fig. 2 and the key structure is shown in equation (2) [22].

$$\mathscr{L}_R = \mathbb{E}_{S,X_1:T\ P}\left[\|S - r_s(e_s(s)\|_2 + \sum_t \|X_t - r_x(e_x(h_s, h_{t-1}, x_t))\|_2\right] \tag{2}$$

where, $\mathscr{L}_R$ is reconstruction loss as our first objective function in the embedding and recovery functions. e_s is an embedding network for static features, e_x is a recurrent embedding network for temporal features, r_s and r_x are recovery networks for static and temporal embeddings.

Feature prediction model

In this study, the effectiveness of data augmentation was validated through the comparisons between the predicted subbase strain by two deep learning models, the TCN-APReLU model and the Attention-Seq2seq model.

TCN-APReLU model

Temporal Convolutional Network (TCN) is a new approach to apply convolutional structures to the field of time-series prediction [23,24]. The architecture of TCN contains dilated causal convolutions with a large perceptual field, where its network structure was described in our previous study [18].

The APReLU developed by Zhao et al. (2021) integrates a well-designed subnetwork as an embedded module for adaptively estimating the multiplication coefficients to be used in different nonlinear transformations [25]. The structure of APReLU is shown in Fig. 3.

Firstly, the feature map is input into ReLU and GAP to calculate a one-dimensional vector to represent the global information of the positive features. At the same time, another one-dimensional vector representing the global information of the negative features is calculated by propagating the input feature map to min (*x, 0*) function and GAP [25]. After that, two one-dimensional vectors are connected and propagated to Full Connect, Batch Normalization, Rectified Linear Unit and Sigmoid in turn. By using ReLU and Sigmoid activation functions, the computational path can provide two levels of nonlinearity when determining the multiplication coefficients. Finally, a specific nonlinear transformation based on the same equation as the parametric ReLU (PReLU) function was employed to get the output feature map. The PReLU function was expressed by equation (3) [25].

$$y = max(x, 0) + \alpha \bullet min(x, 0)\# \tag{3}$$

Since APReLU can be easily implemented in the deep neural network, this study embedded this module into part of the structure of TCN to achieve more flexible non-linear transformations. The improved TCN structure is shown in Fig. 4. In the new ResBlock of TCN, the old activation function was replaced by the APReLU.

Attention-Seq2seq model

The time-series monitored subbase strain contains sufficient information on road structure performance. The prediction of the subbase strain development can help civil engineers better evaluate the service quality and service life.

The attention-sequence to sequence model, like the Attention-Seq2seq model, usually consists of an encoder and a decoder [27–29]. In this study, the Attention-Seq2seq model was employed to predict the next subbase strain sequences sequence from the previous sensor monitoring data. The monitored historical strain data as well as the environmental factor was transformed into the corresponding hidden layers using Bi-LSTM layers in the encoder. The features of the moisture, temperature and some other monitored data were extracted to form

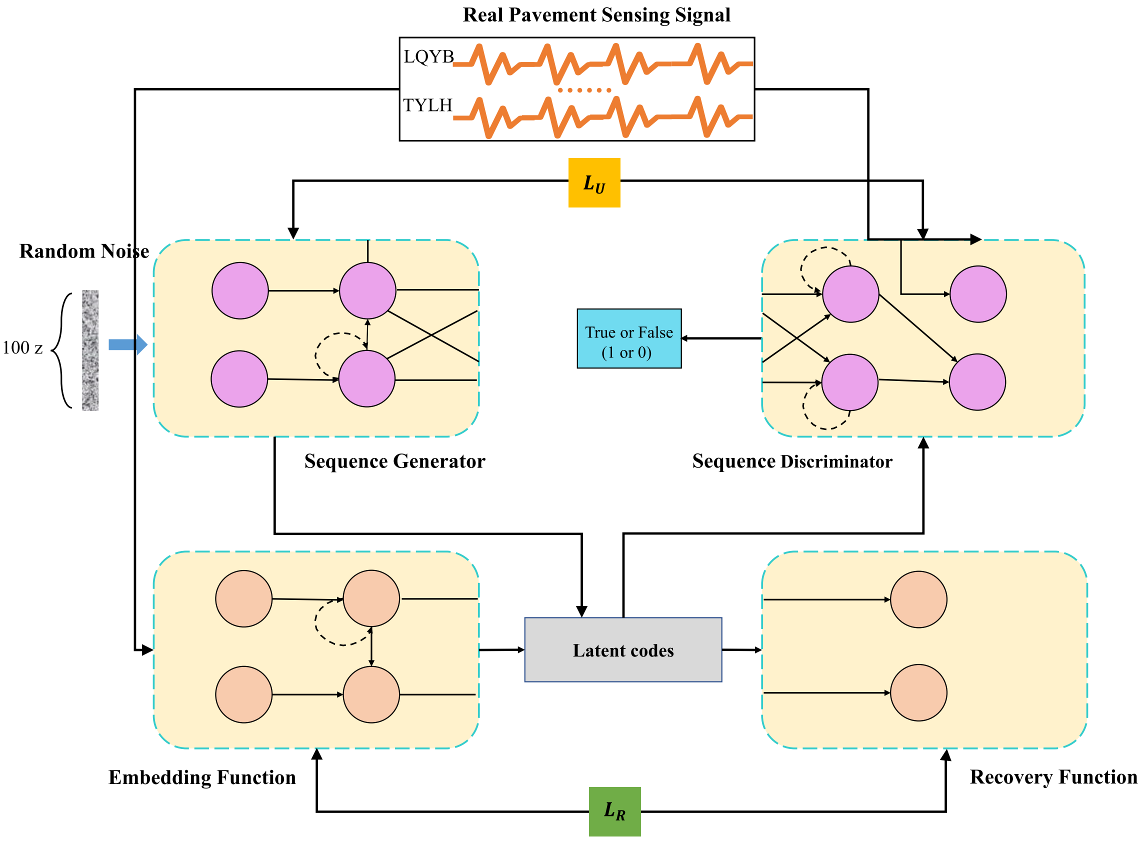


Fig. 2. The architecture of the TimeGAN model [20].

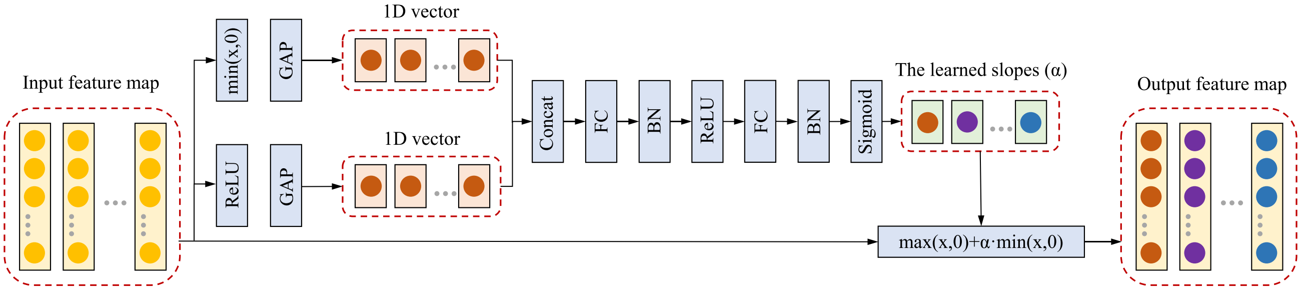


Fig. 3. The structure of the APReLU for adaptive nonlinear transformation [25].

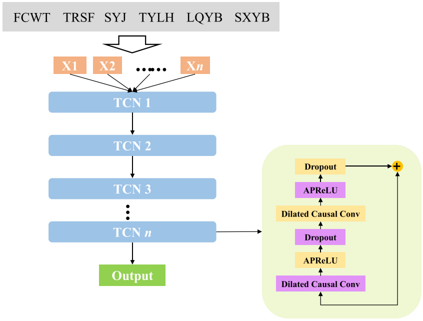


Fig. 4. The structure of the TCN-APReLU [26].

vector *C*, containing the context information from the measured sequence. The context vector was transmitted into the decoder. The own hidden state of the decoder and the current subbase strain were considered to generate the next hidden vector and finally the subbase strain development. The encoder and decoder were calculated following equations (4) and (5) [30].

$$h_t = f_1(X_t, h_{t-1}) \tag{4}$$

$$c = \beta(h_1, \cdots, h_T)$$

and

$$s_t = f_2(Y_{t-1}, s_{t-1}, c) \tag{5}$$

$$p\{Y_t|Y_{t-1}, \cdots, Y_1, c\} = g(Y_{t-1}, s_t, c)$$

where, h_t is the hidden state of encoding Bi-LSTM at time t, s_t is the hidden state of decoding LSTM at time t, c is the last time step of the compressed hidden state from the encoding.

The performance of the Seq2Seq model was related to the length of the input sequence. The shorter input sequence may improve the model performance and less information of the measured data was included in the sequence. In addition, the impact of features varies with different step lengths. An attention mechanism layer was used to solve the problem. The attention layer can make the decoder focus on the most related location, which carried the important information from the encoder sequence. The attention layer can adaptively select as input the hidden states generated by the relevant encoder at all time steps to

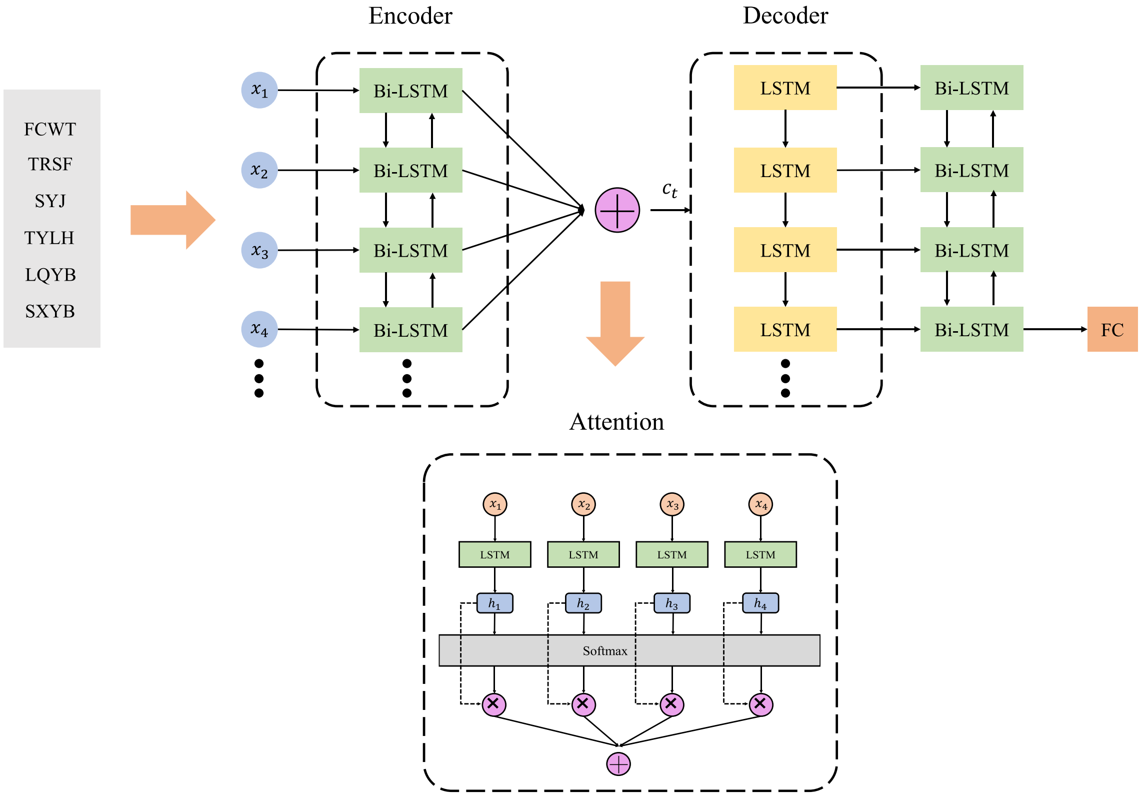


Fig. 5. The structure of the Seq2seq model with attention [27].

predict the future sequence. Fixed context vector c was replaced by variable context vectors c_i calculated via the attention mechanism, as shown in equation (6) [27]. Fig. 5 shows the structure of the Seq2seq model with attention.

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{6}$$

Data preparation

The used data in this study was collected from filed measurements using the buried sensors, which has been documented in previous research [18]. The field measurement was conducted on the Nancun-Shimenying section of the 108 National Highway. The structure information and the location of the sensor was shown in Fig. 6 [18]. As introduced in literature [18], in the figure, the blue line represents the soil pressure sensor (TYLH) in the base layer (ATB-25, anti-fatigue layer, and grade crushed stone layer), the purple line represents the temperature sensor (FCWT) in the entire road structure, the red line represents the asphalt strain sensor (LQYB) in the base layer (ATB-25), the orange line represents the triaxial strain sensor (SXYB) in the subbase layer (grade crushed stone layer), the green line represents the soil moisture sensor (TRSF) and the yellow line represents the seepage pressure sensor (SYJ) in the subgrade layer (improved soil layer). As documented in previous research [18], the strain at the subbase layer had a relationship with the output subbase strain. The collected measured data in this study included the strain at different depths, as well as the temperature and the soil moisture, etc. The measured temperature by the FCWT within each layer was averaged and treated as the layer temperature. The layer temperature of the surface layer (AC-20), base layer (ATB-25, anti-fatigue layer, and grade crushed stone layer), and the subbase, were employed as input temperature time-series. In this study, the input variables of the deep learning model consisted of the vertical pressure collected by TYLH at the bottom of each layer, the layer temperature, the seepage pressure and the moisture of soil, the strain at the bottom of the asphalt layer, and the historical subbase strain. The subbase strain development was predicted using the deep learning model. Generally, the quality of the field measured data was not sufficient to be processed directly due to the environment and traffic factors, and pre-processing of the measured data should be performed first. Also note that this study aims to preliminarily validate the applicability of data-augmentation methods for monitored data prediction, and the corrections of errors generated in the monitoring process from the sensors including soil pressure, soil strain, etc., are not considered. It is also noted that some other important factors, like the strain responses of other structural layers, the layer thickness and material properties, are not considered in the current study due to the limitations of the instruments. In future studies, more mechanical parameters will be monitored and will be considered in comprehensive analysis.

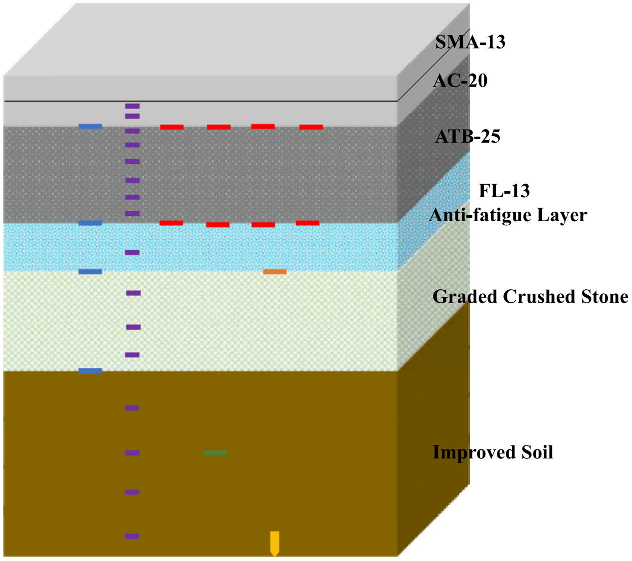


Fig. 6. Sensor placement (Reprinted from Hou Y, Zhao S, Xue Z, Liu S, Song B, Wang D, et al. Intelligent analysis of subbase strain based on a long-term comprehensive monitoring. Transportation Geotechnics, 2022, 33:100720., with permission from Elsevier.) [18].

Data complementation using the Random Forest model

Due to sensor damage, network communication failure, and shortage of power supply during the monitoring process, the data-missing phenomenon occurs during the data collection process and poses a serious concern for data analysis. Current data completion methods include complementary zeros, linear interpolation, mean interpolation, etc. However, these traditional methods may cause deviations from the original data distribution. Random forest is a machine learning algorithm based on decision trees, each of which is independent and can be processed in parallel to enhance the computational efficiency [31,32]. In this study, the random forest was used to predict these missing values and ultimately achieve the purpose of dataset completion. The data complementation based on the random forest algorithm was then performed. Part of the time-series data before the time when missing data occurs was treated as train data and was treated as test data. The features of the train set were set as the target value Y and the features of the remaining set were treated as X. The random forest algorithm was employed to train the data by learning the relationship between the target value Y and X. Then the random forest model was trained to learn the relationship of the target feature Y and X. Finally, the trained model was used to predict the missing data and fill the measured data. The detailed procedure is referred to [33].

The data-missing phenomenon occurred in the measured seepage pressure, the soil moisture, and the temperature data. The data complementation was performed based on the method mentioned above. Fig. 7 shows the effect of the fill on a part of the sensor data. The red curve shows the new data filled by random forest, and the blue curve shows the original data. A reasonable variation trend could be observed in the complemented data, indicating the effectiveness of the data complementation using the RF method.

Data smoothing using Savitzky-Golay (SG) filter method

The noisy signal due to the environmental factor will cause unwanted non-smoothness and the risk of overfitting in the prediction process. This accuracy reduction caused by the noisy signal could be compensated via data smoothing. Therefore, appropriate data smoothing plays a key role in the improvement of subbase strain prediction accuracy. The SG filtering method, which has been considered an efficient approach for the time-series signal process [34,35], was adopted in this study to perform a smoothing operation on the time-series data obtained for each time period.

The core operation in SG filtering includes a convolution operation on time-series data with a specified length, and the curve fitting of the time-series using a polynomial function. The data length in convolution operation was controlled by a window size m. For each convolution window, a polynomial function, as shown in equation (9) [34], was employed to fit the time-series data. It should be noted that the polynomial order K should be less than $2m + 1$.

$$X = a_0 + a_1 i + a_2 i^2 + \cdots + a_k i^k$$

$$= \Rightarrow X_j = \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} C_i s_{i+j}, \frac{m+1}{2} \le j \le n - \frac{m-1}{2} \tag{9}$$

where, m is the filter width, C_i is the convolution factor, a_i is the polynomial parameter determined by least squares. The polynomial order

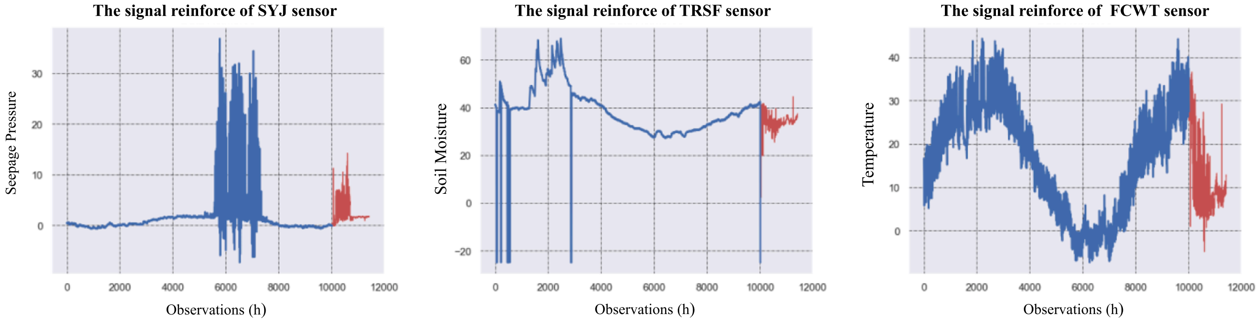


Fig. 7. The effect of the fill on a part of the sensor data.

and the size of window are two serious concerns in GS filtering method including the polynomial order and the size of window. In this study, the window size and polynomial order were determined as 11 and 3 via trial-and-error routine.

Fig. 8 plots the measured strain at the bottom of the asphalt layer and the vertical strain of the triaxial strain after the smoothing operation. It can be observed that the small noises could be removed, and the processed data kept almost same shape and length as the original data. The SG algorithm could effectively improve the smoothness of the original signal and retain the information carried in the measured data.

Results and discussion

Results of the prediction model

In this study, the subbase strain development was predicted through the Seq2Seq and TCN-APReLU models to validate the effectiveness of the data augmentation. The encoder structure of the Seq2seq model was a bidirectional LSTM with 64 cells in the hidden layer, while the decoder structure was an LSTM with 32 cells in the hidden layer. A bidirectional LSTM layer with 64 cells was connected after the encoder-decoder structure, and finally the predicted values were yielded through the fully connected layers. The attention mechanism was a form of

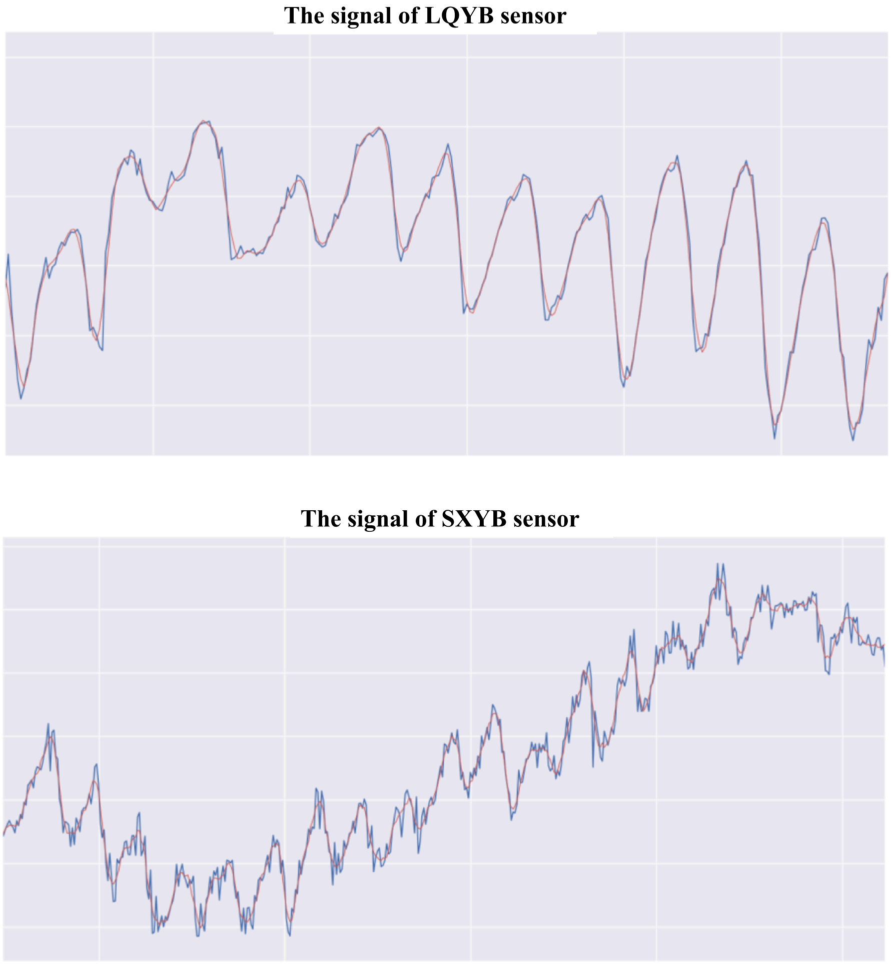


Fig. 8. The smoothing effect of sensor data.

information processing that mimics the distribution of attention in human vision when processing global images, allowing the more critical information to be quickly filtered out from a large amount of useful information to be assigned higher weight values [36,37]. The attention mechanism was shown in Fig. 5. It should be noted that the decoder structure in the Seq2seq model introduced an attention mechanism layer, which enabled the adaptive selection of the relevant encoder hidden state between each time step, thus solving the problem of information loss in fixed vectors in the case of large amounts of information.

The TCN-APReLU model consisted of 13 one-dimensional convolutional layers, which were connected using residuals. The one-dimensional convolutional layers all have a convolutional kernel size of 2, a number of 64, and dilatation factor of 1, 2, 4, 8, 16, and 32 respectively. Each layer used a ReLU activation function followed by a SpatialDropout1D layer with a decay rate set to 0.05, except that the activation function of the last two convolution layers is changed to APReLU.

To quantitatively evaluate the performance of the prediction models, the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) were used as evaluation indices for the prediction performance of each model in this study. The adopted evaluation index was calculated using the following equation shown in (10), (11) and (12).

$$MAE = \frac{1}{N}\sum\nolimits_{i=1}^{N}|y_i - \hat{y}_i| \# \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(y_i - \hat{y}_i)^2 \#} \tag{11}$$

$$MAPE = \frac{100}{N}\sum\nolimits_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \# \tag{12}$$

where, N is the total number of samples in the time-series data, y_i is the real result of the strain sensors monitoring the occurrence of strain in the subbase, and $\hat{y}_i$ is the model-derived prediction of strain in the subbase.

Performance of prediction model based on the original dataset

To evaluate the performance of different prediction models on each segment of the time-series data, the cleaned data was divided into the training set and test set according to the ratio of 8:2. After normalization, the data were input to different deep learning models for training and the number of iterations was set as 150.

Each of the two deep learning models was trained on the training set over nine time periods, and then evaluation indices for all test sets were computed for depolarization averaging and comparing the results of the two models. Table 1 showed the results of each model on the original dataset. It can be seen that the calculated evaluation index from TCN-APReLU was slightly lower than those from Seq2Seq, especially for the RMSE. The convolutional receptive field of the TCN-APReLU can recall more sufficient information of the time-series data. The extraction of static and dynamic features was effectively enhanced through the combination of the residual structure. Then, the gradient disappearance problem can be solved to a certain extent and a better performance can be achieved.

Fig. 9 plotted the prediction results from different deep learning models on time periods. The red curve represents the predicted values, and the blue curve represents the real values. For the data within a longer time period, the agreement was significantly better than that within a short time period. The measured data over a longer time period contained more useful information, which could provide sufficient features for the time-series prediction. It seems that the predicted strain curve from TCN-APReLU matched better than that from Seq2Seq, indicating a better performance of TCN-APReLU. As the monitoring data contains many influencing factors in one sequence, the larger convolutional field of the TCN-APReLU can capture more information for a longer period, and the residual structure was more effective in avoiding information loss in the feature extraction process. It should be noted that there exist considerable differences between the prediction values and the real data, however, the trend of the prediction value is similar to that of the monitored data to a certain extent. Future studies will take this issue into consideration and continuously improve the prediction model accuracy.

Table 1
Predictive performance of models.

Model	Evaluation Metrics		
	MAE	RMSE	MAPE (%)
TCN-APReLU	16.49	24.99	6.28
Attention-Seq2seq	29.40	34.70	9.00

Performance of prediction model based on the augmented dataset

The measured data consisted of the strain at the bottom of each layer, the temperature at different depths, and the soil moisture, as shown in Fig. 6. There was a total of 11 measured sequences to be analyzed in the following section to perform the subbase strain development. The TimeGAN mentioned above was used to augment sensor data with temporal features for nine time periods. The Adam optimizer with a batch size of 128 and sliding window size of 24 was used as training engine. Since the network structure is composed of a self-encoder, generator and discriminator using gate recurrent unit (GRU). The training was split up into 3 main phases, including embedding training, supervised training and joint training, and the number of training iterations was 15,000. The measured data were augmented using the VAE and WGAN-GP for comparison purposes [38,39].

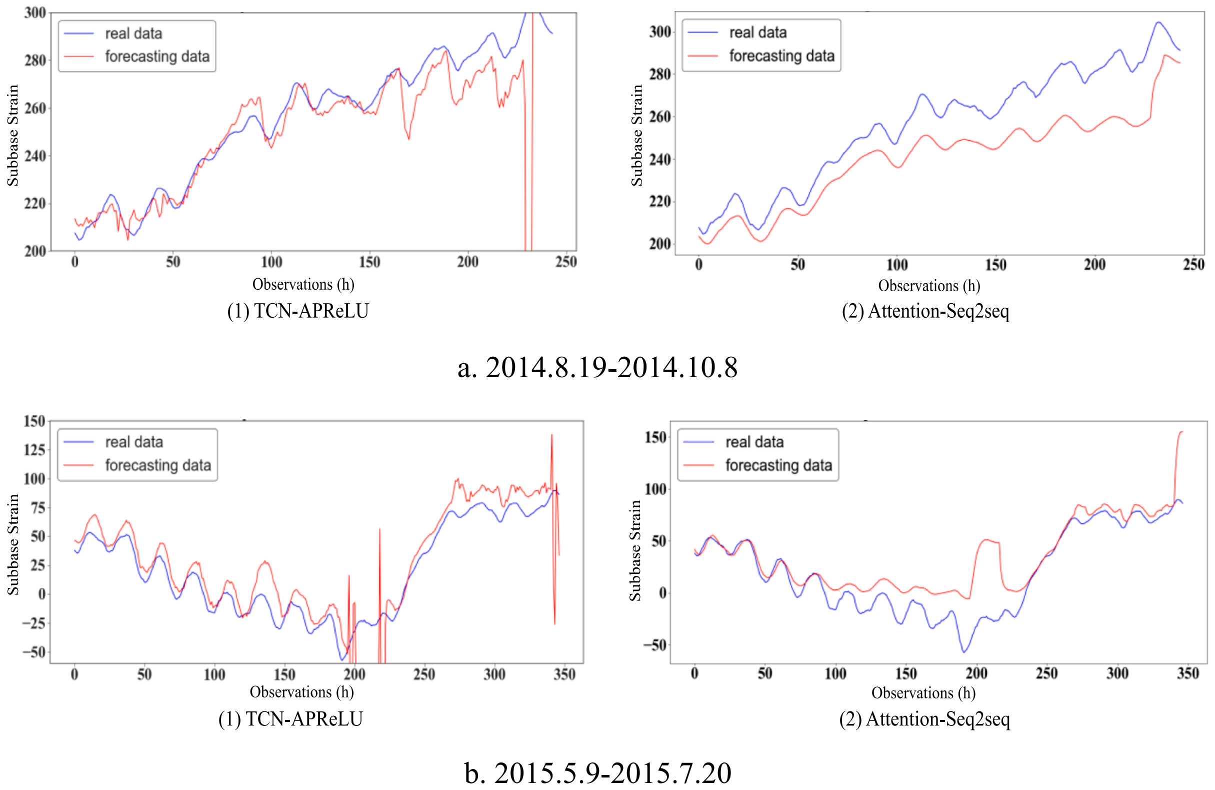
The complexity of the heterogeneous dataset will lead to weak effectiveness of data augmentation application and the generalization ability of data-driven models [40]. Therefore, the data augmentation strategy in this study was the application of a transfer learning algorithm to fuse the augmented data with the actual data. The two deep learning models continuously learned the key temporal information of the augmented synthetic dataset during the training iterations, and then migrated the last updated training weights to their small dataset during the training process. This ensured that the prediction models were trained on the small dataset with prior knowledge, facilitating the learning process to improve the generalization and accuracy of the models.

The length of the augmented data was greater than the original data. More than 50 epochs were performed for the deep learning model training on the augmented data to capture the feature of the augmented time-series data. The model weights and biases were extracted from the training process on the augmented time-series data and transferred directly to the neural network in the target domain, using them as initial parameters for the models trained on the original small data. The weight transfer process is calculated the equation (13) [41].

$$y_t = f\left(W_i^{ori}{}_{(W_i{}^{aug})}x_i + b_i^{ori}{}_{(b_i{}^{aug})}\right) \tag{13}$$

where, W_i^{aug} and b_i^{aug} denote the weights and biases derived from training the model on the augmented dataset; x_i and y_t denote the input and output of each prediction model. The TCN-APReLU and Seq2Seq were employed to train on the augmented data using TimeGAN, VAE and WGAN-GP. The evaluation indices were calculated to evaluate the performance of the prediction method.

Fig. 10 shows the evaluation indices distribution of prediction results on augmented data. The evaluation parameters for the original data were also plotted for comparison. It can be seen that the predicted error of the augmented data from TimeGAN and VAE was less than that of the original data, while the error of the augmented data from WGAN-GP was even greater than that of the original data in some evaluation indices.



(1) TCN-APReLU

(2) Attention-Seq2seq

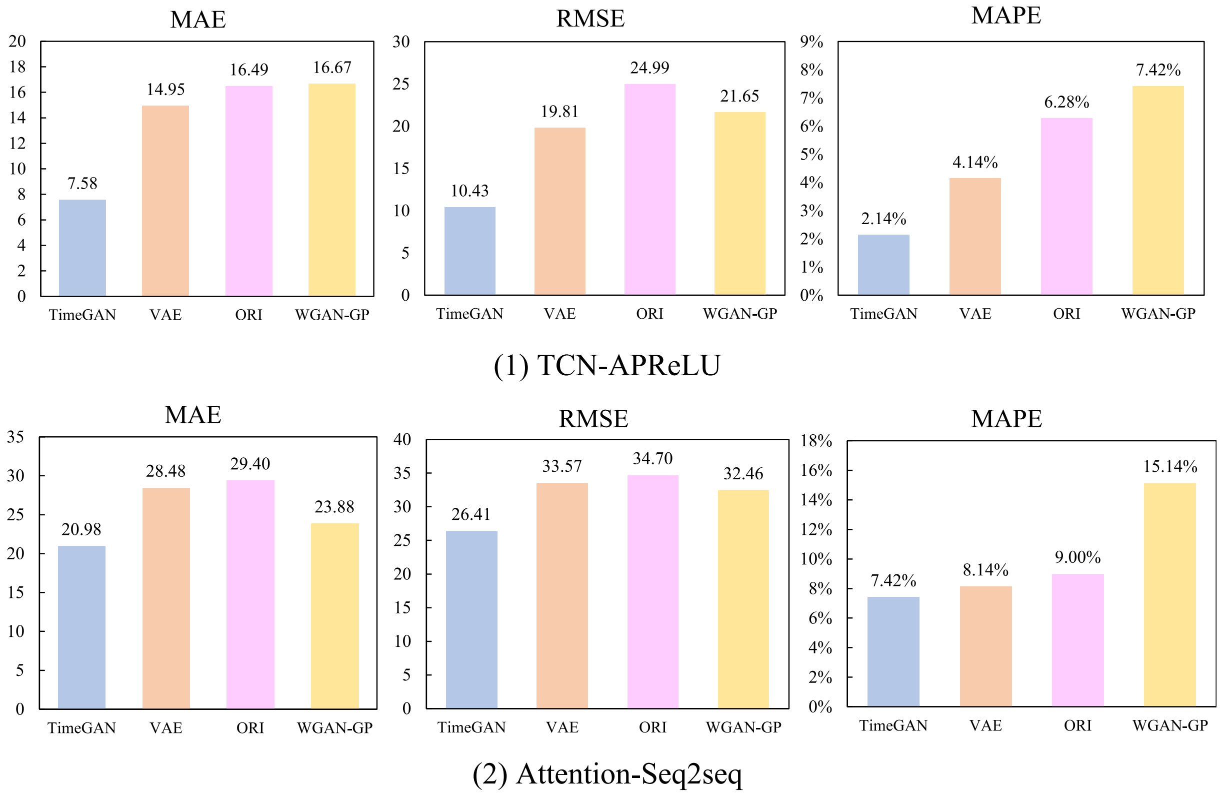
a. 2014.8.19-2014.10.8

(1) TCN-APReLU

(2) Attention-Seq2seq

b. 2015.5.9-2015.7.20

Fig. 9. The prediction results in the test set for different deep learning models.

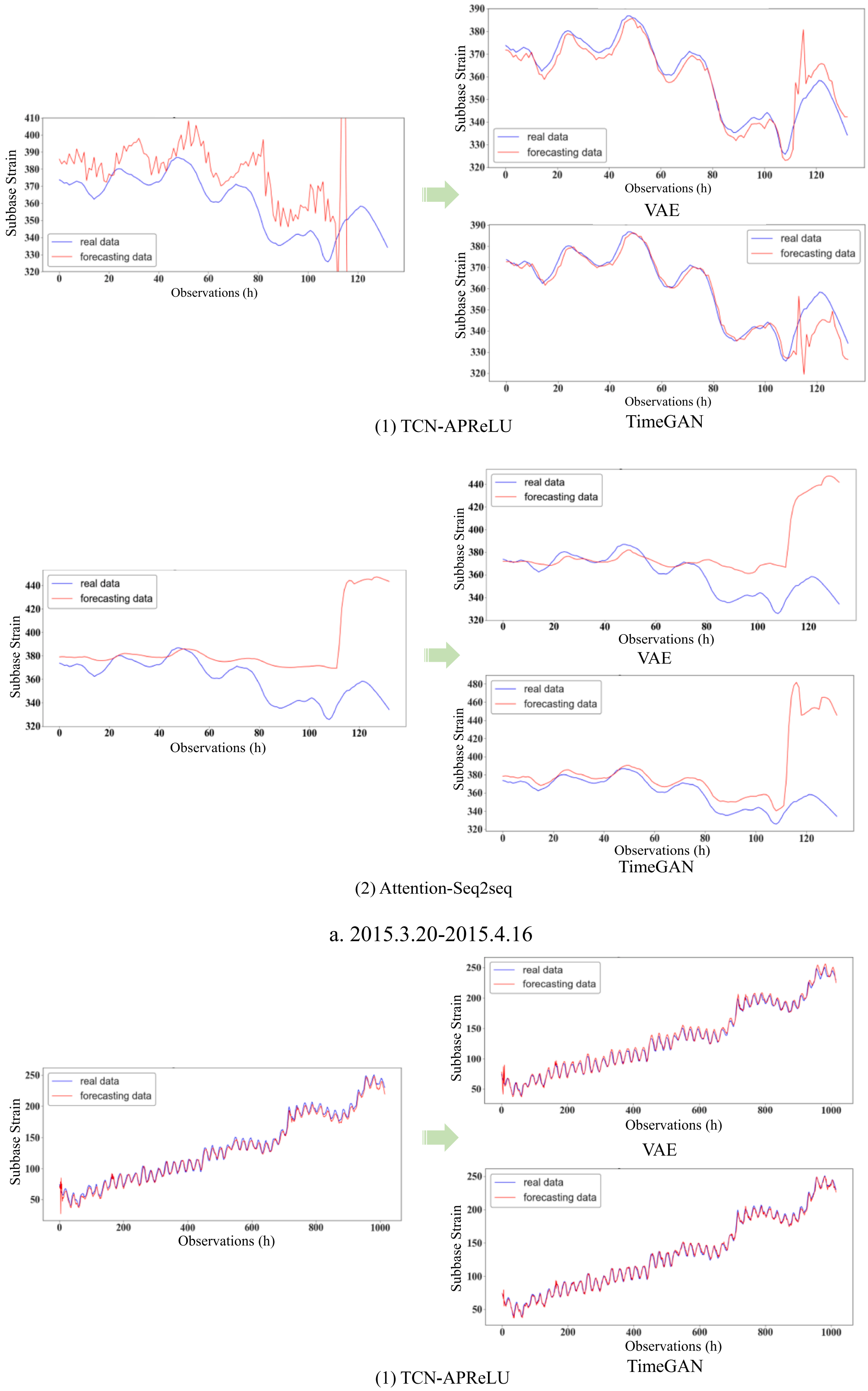


(1) TCN-APReLU

(2) Attention-Seq2seq

Fig. 10. Comparison of evaluation indicators of different augmentation methods.

The result indicated that data augmentation could significantly enhance the prediction performance. The MAE of TCN-APReLU decreased from 16.49 to 7.58 and the RMSE decreased from 24.99 to 10.43, respectively. The MAE and RMSE of Seq2Seq were 20.98 and 26.41. The result denoted that the prediction performance of TCN-APReLU was better than that of Seq2Seq. Among the analyzed case, the prediction performance on the augmented data from TimeGAN was significantly better than other augment method. It can be concluded that TimeGAN captured the temporal dynamic characteristics of the monitoring data well and thus generated high quality synthetic data. The deep network with a residual structure can extract more sufficient temporal features from the augmented data. The significant change of evaluation parameters in analyzed cases reflected the fact that the shallow network with an encoder-decoder structure such as the Attention-Seq2seq was not



VAE

(1) TCN-APReLU

TimeGAN

VAE

TimeGAN

(2) Attention-Seq2seq

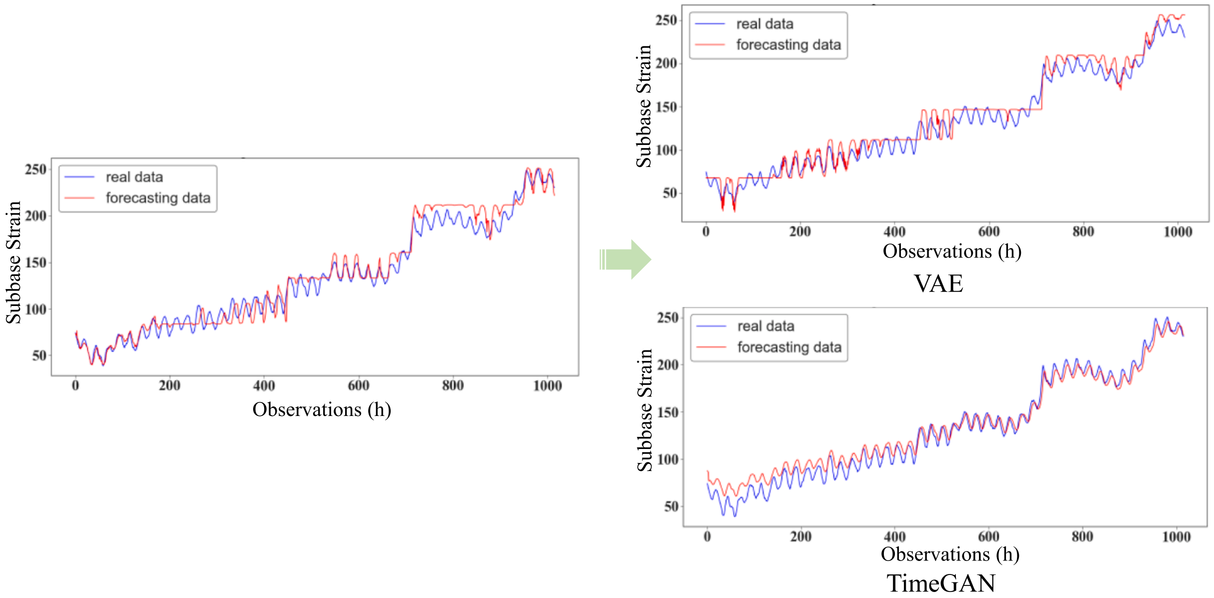
a. 2015.3.20-2015.4.16

VAE

(1) TCN-APReLU

TimeGAN

Fig. 11. Changes in the prediction results of different deep learning models after different data augmentation methods.



VAE

TimeGAN

(2) Attention-Seq2seq

b. 2016.3.4-2016.10.1

Fig. 11. (*continued*).

sensitive to the data amount. This may be due to the weak non-linear representation of the model, which was difficult to fit complex input features with large amounts of data. TCN-APReLU model and TimeGAN exhibited the best prediction performance among the analyzed cases.

The principle of the VAE is to obtain the latent variables through the encoder process and then convert the latent variables into augmented data in the real domain through the decoder process. However, it was clearly seen that the MAE, RMSE and MAPE of TCN-APReLU and Attention-Seq2seq based on VAE only decreased by 1.54, 5.18 and 2.14 %, and 0.92, 1.13 and 0.86 % respectively. This probably occurred because VAE was unable to learn the distribution of the monitoring data accurately. This makes the VAE weak in ability to reconstruct similar data, so the pre-trained model obtained on a large amount of deviated simulated data was equivalent to introducing more noise into the prediction exercise for the real data. The WGAN-GP-based data augmentation method proved to be the worst performer with respect to the predictive performance of the two deep learning models. This phenomenon also corresponds to the fact that the distribution of the real and generated data deviates significantly.

Fig. 11 plotted the typical predicted strain curves on different data augmentation methods of selected periods. The predicted strain on the original data was also plotted on the left side. It can be seen that the agreement of the predicted strain on the augmented data was better than that on the original data. The difference between the predicted strain and the true strain for augmented data was significantly less than that for the original data. The result indicated that the prediction accuracy of the deep learning model had been effectively enhanced by the data augmentation. TimeGAN exhibited better performance for the representation of temporal features than the VAE, and the prediction accuracy could be sequentially enhanced by the TimeGAN. The prediction accuracy of TCN-APReLU was better than Seq2Seq for all the analyzed cases, indicating that the combination of the TimeGAN and TCN-APReLU could provide an effective way to predict the subbase strain development.

Conclusions

A data augmentation-based approach was proposed to predict the subbase strain development in this study. The TimeGAN model was employed to augment the field monitored data for the repair of the leakage of the measurement. The activation function in TCN model was modified for better performance. The correlation between the measured data feature and subbase strain feature was learned through deep learning models to predict the subbase strain development. The effectiveness of the data augmentation on the prediction was evaluated using two deep learning methods. The following conclusions could be drawn from the results.

(1) On the original dataset, the MAPE of TCN-APReLU and Attention-Seq2seq were 6.28 % and 9.00 % respectively, with TCN-APReLU giving the best prediction results, which were also reflected by MAE and RMSE.
(2) The TimeGAN model could effectively augment the measured time-series data, and the augmented data has a good agreement with the monitored data. The TimeGAN could significantly improve the prediction accuracy than VAE and WGAN-GP. Especially for the TCN-APReLU, the MAE for the augmented data was almost 40 % of that for the original data.
(3) The TCN-APReLU model combined with TimeGAN exhibited a good performance for the subbase strain prediction. The TCN-APReLU model could accurately capture the features of the time-series data, and data augmentation could improve the prediction accuracy. The shallow network of Seq2Seq was not sensitive to the data amount, the data augmentation showed a slight influence on the prediction accuracy of Seq2Seq.

Although this study gives a good prediction of subbase strain using a prediction framework based on data augmentation with TimeGAN and TCN-APReLU, however there are still some shortcomings and limitations. In future research work, the following some key points should be focused on: firstly, improvements in the TimeGAN model need to be made in terms of network structure since the original structure of TimeGAN was employed in this study; secondly, advanced temporal prediction algorithms need to be employed, so as to continuously improve the accuracies of predicted values; thirdly, appropriate interpretability analysis of deep learning models to explore the relationships between features needs to be conducted; fourthly, more model evaluation indicators should be used in future studies; finally, the quality of entire monitoring project needs to be improved to include more monitored mechanical parameters and material properties, thus continuously improve the input of the model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgement

This work was supported by the Opening project fund of Materials Service Safety Assessment Facilities (MSAF- 2021-109), the International Research Cooperation Seed Fund of Beijing University of Technology (No. 2021A05), the National Natural Science Foundation of China (grant number 52008012), and Hunan Expressway Group Co. Ltd and the Hunan Department of Transportation (No. 202152) in China. The authors would also like to give sincere gratitude to Dr. Dandan Cao for his valuable work on the language editing of this manuscript.

References

[1] Lu C, Chen J, Gu C, Wang J, Cai Y, Zhang T, et al. Resilient and permanent deformation behaviors of construction and demolition wastes in unbound pavement base and subbase applications. Transp Geotech 2021;28:100541. https://doi.org/10.1016/j.trgeo.2021.100541.

[2] Ma X, Dong Z, Yu X, Chen F, Cao C, Sun J. Monitoring the structural capacity of airfield pavement with built-in sensors and modulus back-calculation algorithm. Constr Build Mater 2018;175:552–61. https://doi.org/10.1016/j.conbuildmat.2018.04.198.

[3] Xin X, Liang M, Yao Z, Su L, Zhang J, Li P, et al. Self-sensing behavior and mechanical properties of carbon nanotubes/epoxy resin composite for asphalt pavement strain monitoring. Constr Build Mater 2020;257:119404. https://doi.org/10.1016/j.conbuildmat.2020.119404.

[4] Xue W, Wang L, Wang D, Druta C. Pavement Health Monitoring System Based on an Embedded Sensing Network. J Mater Civ Eng 2014;26:04014072. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000976.

[5] Hou Y, Li Q, Zhang C, Lu G, Ye Z, Chen Y, et al. The State-of-the-Art Review on Applications of Intrusive Sensing, Image Processing Techniques, and Machine Learning Methods in Pavement Monitoring and Analysis. Engineering 2021;7: 845–56. https://doi.org/10.1016/j.eng.2020.07.030.

[6] Bansal T, Talakokula V, Mathiyazhagan K. Equivalent structural parameters based non-destructive prediction of sustainable concrete strength using machine learning models via piezo sensor. Measurement 2022;187:110202. https://doi.org/10.1016/j.measurement.2021.110202.

[7] Tabrizi SE, Xiao K, Van Griensven TJ, Saad M, Farghaly H, Yang SX, et al. Hourly road pavement surface temperature forecasting using deep learning models. J Hydrol 2021;603:126877. https://doi.org/10.1016/j.jhydrol.2021.126877.

[8] Zeng C, Huang J, Xie J, Zhang B, Indraratna B. Prediction of mud pumping in railway track using in-service train data. Transp Geotech 2021;31:100651. https://doi.org/10.1016/j.trgeo.2021.100651.

[9] Zheng Q, Hou Y, Yang H, Tan P, Shi H, Xu Z, et al. Towards a sustainable monitoring: A self-powered smart transportation infrastructure skin. Nano Energy 2022;98:107245. https://doi.org/10.1016/j.nanoen.2022.107245.

[10] Bandara K, Hewamalage H, Liu Y-H, Kang Y, Bergmeir C. Improving the accuracy of global forecasting models using time series data augmentation. Pattern Recogn 2021;120:108148. https://doi.org/10.1016/j.patcog.2021.108148.

[11] Jiang L, Wang Y, Tang Z, Miao Y, Chen S. Casting defect detection in X-ray images using convolutional neural networks and attention-guided data augmentation. Measurement 2021;170:108736. https://doi.org/10.1016/j.measurement.2020.108736.

[12] Huang H, Zhou H, Yang X, Zhang L, Qi L, Zang A-Y. Faster R-CNN for marine organisms detection and recognition using data augmentation. Neurocomputing 2019;337:372–84. https://doi.org/10.1016/j.neucom.2019.01.084.

[13] Wu H, Liu H. Non-intrusive load transient identification based on multivariate LSTM neural network and time series data augmentation. Sustainable Energy Grids Networks 2021;27:100490. https://doi.org/10.1016/j.segan.2021.100490.

[14] Jeong CY, Shin HC, Kim M. Sensor-data augmentation for human activity recognition with time-warping and data masking. Multimed Tools Appl 2021;80: 20991–1009. https://doi.org/10.1007/s11042-021-10600-0.

[15] Almonacid F, Pérez-Higueras P, Rodrigo P, Hontoria L. Generation of ambient temperature hourly time series for some Spanish locations by artificial neural networks. Renew Energy 2013;51:285–91. https://doi.org/10.1016/j.renene.2012.09.022.

[16] Taylor D. Using a multi-head, convolutional neural network with data augmentation to improve electropherogram classification performance. Forensic Sci Int Genet 2022;56:102605. https://doi.org/10.1016/j.fsigen.2021.102605.

[17] Fu R, Chen J, Zeng S, Zhuang Y, Sudjianto A. Time Series Simulation by Conditional Generative Adversarial Net. ArXiv:190411419 [Cs, Eess, Stat] 2019.

[18] Hou Y, Zhao S, Xue Z, Liu S, Song B, Wang D, et al. Intelligent analysis of subbase strain based on a long-term comprehensive monitoring. Transp Geotech 2022;33: 100720. https://doi.org/10.1016/j.trgeo.2022.100720.

[19] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in Neural Information Processing Systems 27 (nips 2014), vol. 27, La Jolla: Neural Information Processing Systems (nips); 2014, p. 2672–80.

[20] Baasch G, Rousseau G, Evins R. A Conditional Generative adversarial Network for energy use in multiple buildings using scarce data. Energy and AI 2021;5:100087. https://doi.org/10.1016/j.egyai.2021.100087.

[21] Öcal A, Özbakır L. Supervised deep convolutional generative adversarial networks. Neurocomputing 2021;449:389–98. https://doi.org/10.1016/j.neucom.2021.03.125.

[22] Yoon J, Jarrett D, van der Schaar M. Time-series Generative Adversarial Networks. Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc.; 2019.

[23] Dai W, An Y, Long W. Price change prediction of Ultra high frequency financial data based on temporal convolutional network. Procedia Comput Sci 2022;199: 1177–83. https://doi.org/10.1016/j.procs.2022.01.149.

[24] Wang J-J, Wang C, Fan J-S, Mo YL. A deep learning framework for constitutive modeling based on temporal convolutional network. J Comput Phys 2022;449: 110784. https://doi.org/10.1016/j.jcp.2021.110784.

[25] Zhao M, Zhong S, Fu X, Tang B, Dong S, Pecht M. Deep Residual Networks With Adaptively Parametric Rectifier Linear Units for Fault Diagnosis. IEEE Trans Ind Electron 2021;68:2587–97. https://doi.org/10.1109/TIE.2020.2972458.

[26] Hewage P, Behera A, Trovati M, Pereira E, Ghahremani M, Palmieri F, et al. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput 2020;24: 16453–82. https://doi.org/10.1007/s00500-020-04954-0.

[27] Liang Z, Du J, Li C. Abstractive social media text summarization using selective reinforced Seq2Seq attention model. Neurocomputing 2020;410:432–40. https://doi.org/10.1016/j.neucom.2020.04.137.

[28] Ma Z, Du B, Shen J, Yang R, Wan J. An Encoding Mechanism for Seq2Seq based Multi-Turn Sentimental Dialogue Generation Model. Procedia Comput Sci 2020; 174:412–8. https://doi.org/10.1016/j.procs.2020.06.108.

[29] Sun K, Qian T, Chen X, Zhong M. Context-aware seq2seq translation model for sequential recommendation. Inf Sci 2021;581:60–72. https://doi.org/10.1016/j.ins.2021.09.001.

[30] Zhang Y, Li Y, Zhang G. Short-term wind power forecasting approach based on Seq2Seq model using NWP data. Energy 2020;213:118371. https://doi.org/10.1016/j.energy.2020.118371.

[31] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[32] Wang J, Song Z, Chen L, Xu T, Deng L, Qi Z. Prediction of CO2 solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors. Green Chemical Engineering 2021. https://doi.org/10.1016/j.gce.2021.08.002.

[33] Li J, Wang Z, Lai C, Zhang Z. Tree-ring-width based streamflow reconstruction based on the random forest algorithm for the source region of the Yangtze River. China CATENA 2019;183:104216. https://doi.org/10.1016/j.catena.2019.104216.

[34] Wu J-M-T, Tsai M-H, Huang YZ, Islam SH, Hassan MM, Alelaiwi A, et al. Applying an ensemble convolutional neural network with Savitzky-Golay filter to construct a phonocardiogram prediction model. Appl Soft Comput 2019;78:29–40. https://doi.org/10.1016/j.asoc.2019.01.019.

[35] Zhang G, Hao H, Wang Y, Jiang Y, Shi J, Yu J, et al. Optimized adaptive Savitzky-Golay filtering algorithm based on deep learning network for absorption spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc 2021;263:120187. https://doi.org/10.1016/j.saa.2021.120187.

[36] Alizadeh B, Ghaderi Bafti A, Kamangir H, Zhang Y, Wright DB, Franz KJ. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. J Hydrol 2021;601:126526. https://doi.org/10.1016/j.jhydrol.2021.126526.

[37] Wang Y, Zhang X, Lu M, Wang H, Choe Y. Attention augmentation with multi-residual in bidirectional LSTM. Neurocomputing 2020;385:340–7. https://doi.org/10.1016/j.neucom.2019.10.068.

[38] Moreno-Barea FJ, Jerez JM, Franco L. Improving classification accuracy using data augmentation on small data sets. Expert Syst Appl 2020;161:113696. https://doi.org/10.1016/j.eswa.2020.113696.

[39] Gao X, Deng F, Yue X. Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty.

Neurocomputing 2020;396:487–94. https://doi.org/10.1016/j.neucom.2018.10.109.

[40] Lu Y, Tian Z, Zhang Q, Zhou R, Chu C. Data augmentation strategy for short-term heating load prediction model of residential building. Energy 2021;235:121328. https://doi.org/10.1016/j.energy.2021.121328.

[41] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33. https://doi.org/10.1007/BF02478259.