IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2021, accepted July 3, 2021, date of publication July 15, 2021, date of current version July 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3097353

Drought Prediction Based on Feature-Based Transfer Learning and Time Series Imaging

WAN TIAN, JIUJING WU, HENGJIAN CUI, AND TAO HU
School of Mathematical Sciences, Capital Normal University, Beijing 100048, China

Corresponding author: Tao Hu (hutaomath@foxmail.com)

This work was supported in part by the Key Projects of the National Natural Science Foundation of China under Grant 12031016, in part by the National Natural Science Foundation of China under Grant 11971324 and Grant 11471223, in part by the Interdisciplinary Construction of Bioinformatics and Statistics, in part by the Academy for Multidisciplinary Studies, Capital Normal University, and in part by the Beijing Science and Technology Innovation Platform Construction Project.

ABSTRACT Drought is an extreme climate phenomenon that has a great impact on the economy, tourism, agriculture, and water resources. Drought prediction can provide an early warning of the occurrence of drought and reduce losses. In this article, the standard precipitation evapotranspiration index (SPEI) on four time scales: SPEI-3, SPEI-6, SPEI-9, and SPEI-12 are used to measure and predict drought. Unlike the general methods of directly modeling the SPEI index, time-series imaging and feature-based transfer learning are used to extract the features of the SPEI sequence and use the extracted features for prediction. First, we use Gramian Angular Summation/Difference Field (GASF/GADF), Markov Transition Field (MTF), and Recurrence Plot (RP) as the time series imaging techniques to encode SPEI sequences into images. Secondly, we utilize imaging data sets and convolutional neural networks (CNNs) such as residual network (ResNet) and VGG to train the feature extraction network. Finally, the following four regressors: Random Forest (RF), Long and Short-Term Memory network (LSTM), Wavelet Neural Network (WNN), Support Vector Regression (SVR) are used to model the extracted features and drought prediction. To verify the effectiveness of the method proposed in this article, we use the SPEI of four time scales at eight stations in the Haihe River Basin for prediction. Compared with the existing methods, the prediction results of different time scales and stations are improved. For example, after feature extraction, LSTM can reach MAPE $= 0.5400$, SMAPE $= 0.4452$, MAE $= 0.2150$, MSE $= 0.0853$ and $R^2 = 0.8960$ in the SPEI-12 prediction of the Beijing site, and other results show that the proposed method is not sensitive to the time scale of drought prediction.

INDEX TERMS Drought prediction, deep learning, imaging, transfer learning.

I. INTRODUCTION

Drought refers to the periodical precipitation on land that is lower than normal for many months or years [8]. Drought can be divided into meteorology drought, agriculture drought, hydrology, and socio-economic drought [22]. Drought causes huge economic, political and cultural losses every year. Therefore, it is crucial to predict drought and take corresponding measures. Many models have been developed for this purpose. An excellent drought prediction model is of great significance to the planning and management of water resources and minimizing the negative effects of drought. In practice, droughts can also be classified based on time scales and precipitation anomalies, such as the standardized precipitation index (SPI) [20] and standardized precipitation evaporation index (SPEI) [20] based on different time scales that are often used to measure drought levels: SPEI-1 to SPEI-24 [12]. There are many types of research on drought prediction, including drought evaluation indicators, models, research areas, etc. There is no doubt that the drought prediction model is the most important and basic. According to the mechanism, the drought prediction model can be divided into regression analysis, stochastic, machine learning-based, hybrids, and dynamic models [4], [12], [13], [28], [30], [39]. Different models are suitable for different research areas, different precipitation conditions, and drought evaluation indicators on time scales.

Essentially, the linear model is the most basic, simplest, and most explanatory model. Logistic regression [21], [28], [31] and log-linear regression [19] in linear model

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

are widely used in drought prediction. [28] applied logistic regression to predict SPI and SPEI in Europe and used the pseudo-correlation coefficient and the area under the receiver operating characteristic curve (AUROC) as evaluation indicators. [21] applied a logistic regression model to analyze the impact of SPI and Southern Oscillation Index (SOI) in East China on seasonal drought in different regions and different seasons. The main advantages of the regression model are simple and direct and low computational cost, but due to the linear assumption, its long-term prediction performance is poor [12]. [24] proposed the use of machine learning-based methods, including RF, SVR, and boosted regression trees (BRT) for drought prediction in the United States. The most important and popular stochastic models are ARIMA [23], SARIMA [2] and their variants. This type of model has several important parameters, namely autoregressive order p, difference order d and moving average order q. Once the order of the model is confirmed, the ARIMA and SARIMA models can be described as ARIMA(p, d, q) and ARIMA$(p, d, q)(P, D, Q)_s$ respectively, where (p, d, q) and $(P, D, Q)_s$ represent non-seasonal and seasonal parts, respectively. [2], [23] used ARIMA and SARIMA models for hydrological drought prediction in Ethiopia and determined that $(0, 1, 1)(0, 1, 1)_{12}$ was the best model among the candidate models. [37] used multiple models for drought prediction, including WNN, ANN, etc., and achieved good prediction performance at eight stations in the Haihe River Basin. The prediction performance of most of the models proposed above is strongly dependent on the study area, observation sites, drought evaluation indicators, and time scales, and most models directly predict the original data and cannot extract important features from the data.

To fix problems of existing models, we propose a novel feature extraction method based on time series imaging [14] and feature-based transfer learning that can effectively extract drought data features, and its prediction performance does not depend on the study area. In addition, the method proposed in the article can also be combined with any existing prediction model for drought prediction. At present, deep neural networks have reached the state of the art in various tasks, such as computer vision (CV) [15], natural language processing (NLP) [7], [33], and recommendation systems (RS) [9]. One of the main reasons for the success of deep learning (DL) is that DL models can represent the raw data well. Naturally, we use the images obtained by each imaging scheme as a class for training CNNs, also called feature extraction networks, and use to extract features for drought prediction. The CNNs architectures selected in the article are ResNet [15], VGG [26], and their variants. Finally, we use the extracted features for drought prediction. In this article, we use some regressors that perform well in drought prediction, including LSTM [16], RF [6], WNN [3] and SVR [27]. To evaluate the performance of the proposed method, we use common indicators for evaluating prediction performance, including Mean Square Error (MSE), coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE).

The structure of the article is as follows. In section II, we introduce the study area, SPEI, and evaluation indicators of the model. In section III, we introduce the drought prediction methods proposed in the article, including model ordering, time series imaging, and feature extraction based on transfer learning. In section IV, we introduce in detail several regressors used in the article, including LSTM, RF, WNN, and SVR. In section V, we apply the proposed method and model to SPEI prediction and discuss the experimental results.

II. PRELIMINARIES

In this section, we will introduce in detail the study area, the SPEI for evaluating drought, and the index for evaluating model performance that will be used later in the article.

A. STUDY AREA

The Haihe River Basin borders the Bohai Sea in the east, Taihang in the west, the Yellow River in the south, and the Mongolian Plateau in the north. The total area of the basin 3.182×10^5km^2, accounting for 3.3% of the China's total area. The Haihe River Basin includes 3 major river systems and 10 backbone rivers. The total terrain of the whole basin is high in the northwest and low in the southeast, roughly divided into three types of landforms: plateau, mountain and plain. The western part is the Loess Plateau and the Taihang Mountains. The northern part is the Mongolian Plateau, and the Yanshan Mountains, covering an area of 1.894×10^5km^2, accounting for 60%; the east and southeast are plains with an area of 1.284×10^5km^2, accounting for 40% [38].

The basin belongs to the temperate East Asian monsoon climate zone. The winter is controlled by the Siberian continental air mass, which is cold and less snowy; in the spring, it is affected by the Mongolian continental air mass. The temperature rises quickly and the evaporation is large, often forming dry weather; the summer is affected by the marine air mass, which is relatively humid, with more rainfall, and drought occurs; autumn is the transitional season between summer and winter, and the general year is high and cool, with less rainfall.

In this paper, the SPEI with a time scale of 3, 6, 9, and 12 months (SPEI-3, SPEI-6, SPEI-9, SPEI-12) is used to measure drought level and drought prediction. The data used comes from the China Meteorological Administration, which records precipitation data in the Haihe River Basin from 1960 to 2010. The article studied 8 sites in the northern Haihe River Basin, namely Datong, Yuxian, Fengning, Zhangjiakou, Huailai, Zunhua, Beijing, and Tangshan. The basic description of these sites is shown in Table 1.

B. STANDARD PRECIPITATION EVAPORATION INDEX (SPEI)

SPEI is used to describe the difference between weekly or monthly precipitation and potential evapotranspiration (PEP) [20]. The SPEI can reflect the climate water balance

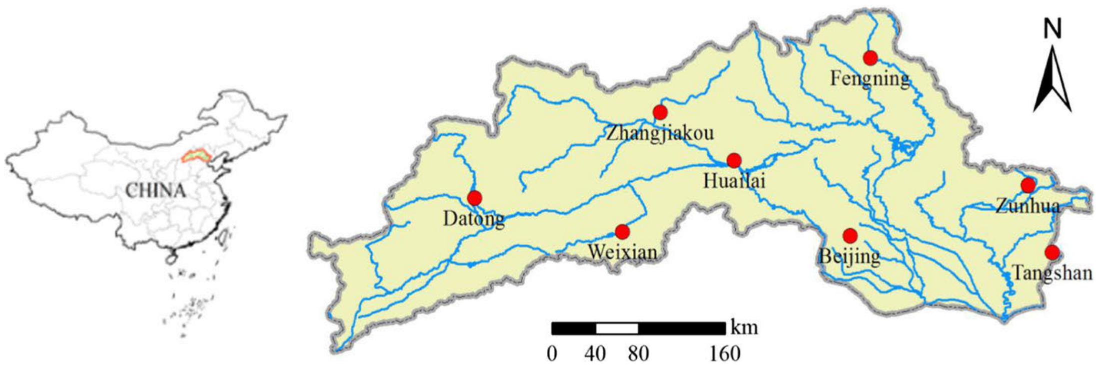

FIGURE 1. Study area and meteorological stations.

TABLE 1. Basic description of the research area.

Site Name	Site Code	Lat(N)	Long(E)	Elevation(m)	Mean AP[1](mm)	Max AP (mm)
Datong	53487	40.1	113.3	1067.2	370.24	579.0
Weixian	53593	39.8	114.6	909.5	398.88	616.3
Fengning	54308	41.2	116.6	661.2	457.90	696.4
Zhangjiakou	54401	40.8	114.9	724.2	399.00	591.5
Huailai	54405	40.4	115.5	536.8	378.98	543.6
Zunhua	54429	40.2	118.0	54.9	711.88	1193.4
Beijing	54511	39.8	116.5	31.3	549.1843	913.2
Tangshan	54534	39.7	118.2	27.8	605.37	1007.7

[1] AP: Annual Precipitation

on different time scales to a certain extent, and can also be used to measure and monitor the degree of drought. Given the precipitation P_i and PET_i in the i-th month, the difference between the two can be calculated by equation

$$D_i = P_i - \mathrm{PET}_i.$$

Generally speaking, climate data such as daily rainfall and streamflow obey the log-logistic distribution, and its probability density function (PDF) is

$$f(z) = \frac{\beta}{\alpha}\left(\frac{z-\gamma}{\alpha}\right)^{\beta-1}\left(1+\left(\frac{z-\gamma^{\beta}}{\alpha}\right)\right)^{-2},\ z \geq \gamma, \tag{1}$$

where α, β, and γ are the scale, shape and location parameters respectively. Obviously, by replacing z in equation (1) with D_i, we can get $\infty > D_i \geq \gamma$. The three parameters of the log-logistic distribution are estimated using

$$\widehat{\alpha} = \frac{(w_0 - 2w_1)\widehat{\beta}}{\Gamma\left(1+\frac{1}{\widehat{\beta}}\right)\Gamma\left(1-\frac{1}{\widehat{\beta}}\right)}, \quad \widehat{\beta} = \frac{2w_1 - w_0}{6w_1 - w_0 - 6w_2},$$
$$\widehat{\gamma} = w_0 - \widehat{\alpha}\Gamma\left(1+\frac{1}{\widehat{\beta}}\right)\Gamma\left(1-\frac{1}{\widehat{\beta}}\right),$$

where $\Gamma(\cdot)$ is gamma function and w_0, w_1, w_2 are the probability weighted moments. After estimating the three parameters, we can get the cumulative distribution function (CDF) of the log-logistic distribution

$$F(z) = \left[1+\left(\frac{\widehat{\alpha}}{z-\widehat{\gamma}}\right)^{\widehat{\beta}}\right]^{-1}. \tag{2}$$

TABLE 2. Correspondence between SPEI and drought level.

SPEI Value	Drough level
$\mathrm{SPEI} > 2.0$	Extremely wet
$1.5 < \mathrm{SPEI} \leq 2.0$	Very wet
$1.0 < \mathrm{SPEI} \leq 1.5$	Moderately wet
$-1.0 < \mathrm{SPEI} \leq 1.0$	Near normal
$-1.5 < \mathrm{SPEI} \leq -1.0$	Moderately dry
$\mathrm{SPEI} \leq -2.0$	Extremely dry

By using the Abramowitz and Stegun approximations to equation (2), transforming it into a standard normal distribution, we can use the following formula to calculate SPEI

$$\mathrm{SPEI} = W - \frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3},$$

where C_0, C_1, C_2, d_1, d_2, d_3 are all constants [1], and W is defined as follows

$$W = \begin{cases} \sqrt{-2\ln(P)}, & \text{if } P \leq 0.5, \\ \sqrt{-2\ln(P-1)}, & \text{if } P > 0.5, \end{cases}$$

where $P = 1 - F(z)$. Table 2 [37]shows the drought levels corresponding to different SPEI ranges.

C. EVALUATION INDEX

To compare the pros and cons of different forecasting methods, the following five indicators are selected to evaluate the model, including MSE, MAE, MAPE, SMAPE, and R^2. If there is a set of data $y_1, y_2, \cdots, y_t$, its corresponding predicted value is $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_t$, then the corresponding

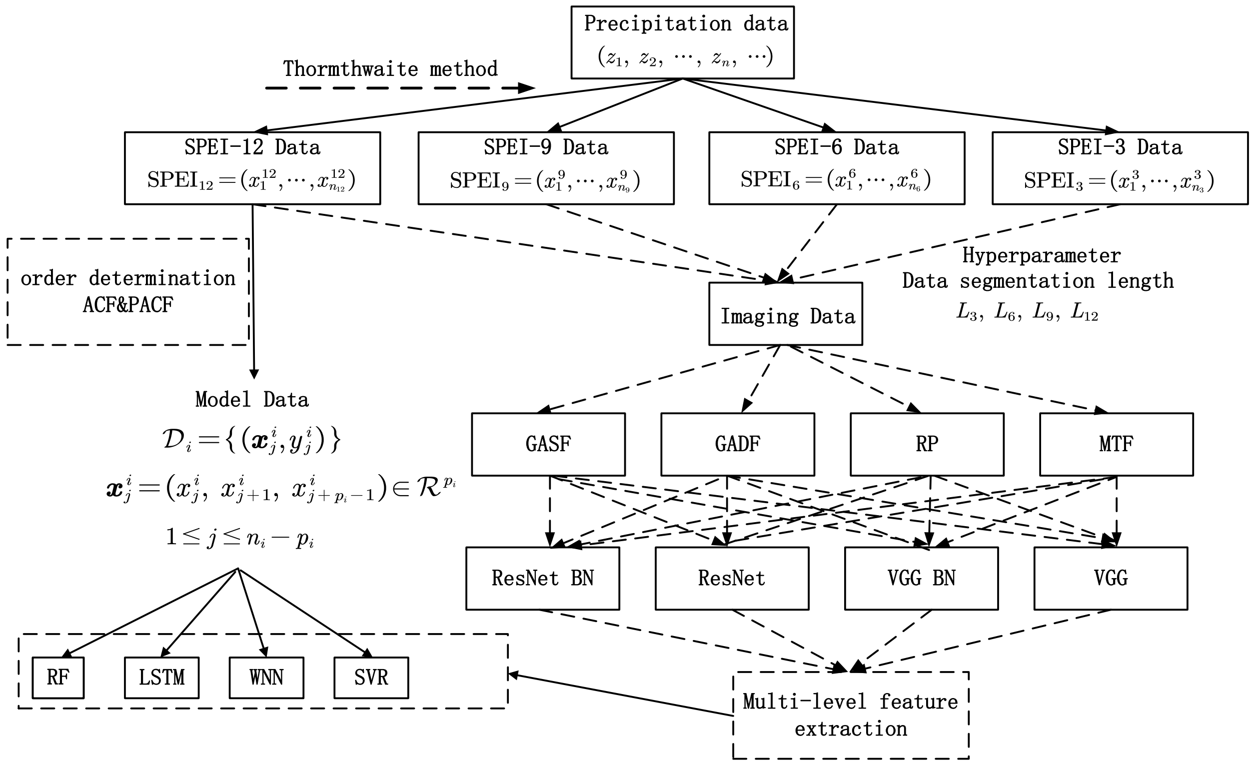

FIGURE 2. Technical roadmap of drought prediction based on time series imaging and transfer learning.

expression of the above evaluation index is

$$\mathrm{MSE} = \frac{1}{t}\sum_{i=1}^{t}(\widehat{y}_i - y_i)^2,$$

$$\mathrm{SMAPE} = \frac{1}{t}\sum_{i=1}^{t}\frac{|\widehat{y}_i - y_i|}{(|y_i| + |\widehat{y}_i|)/2} \times 100\%,$$

$$\mathrm{MAPE} = \frac{1}{t}\sum_{i=1}^{t}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \times 100\%,\ \mathrm{MAE}=\frac{1}{t}\sum_{i=1}^{t}|\widehat{y}_i - y_i|.$$

III. PROPOSED APPROACH

In this part, the methods proposed in this article will be introduced in detail, including the use of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for model ordering, four imaging schemes for sequence data, feature extraction method based on transfer learning and drought prediction models. Let the precipitation data be a series $z = \{z_1, z_2, \cdots, z_n, \cdots\}$. First, according to the formula of SPEI, we convert the precipitation data into four time scales drought indicators SPEI-3, SPEI-6, SPEI-9, SPEI-12, where $\mathrm{SPEI}_i = \left(x_1^i, \cdots, x_{n_i}^i\right), i = 3, 6, 9, 12$. Secondly, we use the ACF and PACF commonly used in time series to determine the order of the sequence SPEI_i. Let the order of SPEI_i sequences be p_i, we can construct data sets

$$\mathcal{D}_i = \left\{\left(\boldsymbol{x}_j^i, y_j^i\right)\right\}, \quad \boldsymbol{x}_j^i = \left(x_j^i, x_{j+1}^i, \quad x_{j+p_i-1}^i\right) \in \mathcal{R}^{p_i}, \quad (3)$$

where $1 \leq j \leq n_i - p_i,\ y_j = x_{j+p_i}^i$. Next, we will use four imaging technologies GADF, GASF, MTF and RP to encode the data $\boldsymbol{x}_j^i$ into an image, and we can get the corresponding image data set. Figure 2 shows the technical roadmap of the method proposed in the article.

A. ORDER DETERMINATION

Because the order of the model is applied to precipitation data, the symbol z is used to represent precipitation series data. Similarly, the four imaging schemes described later will be applied to the SPEI sequence data, so the symbol x will be used. Given a time series, the order of the model determines how many past observations are used for prediction. In this article, ACF and PACF are used to determine the order of the model [11]. Let $\{z_t\}, t = 1, 2, \cdots$ be a time series, the autocovariance function (ACVF) of $\{z_t\}$ is defined as

$$\gamma(k) = \mathrm{Cov}\left(z_{t+k}, z_t\right), \quad k = 0, \pm 1, \cdots. \quad (4)$$

The ACF of $\{z_t\}$ is

$$\rho(k) = \gamma(k)/\gamma(0) = \mathrm{Corr}\left(z_{t+k}, z_t\right), \quad k = 0, \pm 1, \cdots. \quad (5)$$

From equation (4) and (5) we can see that both ACVF and ACF are even functions

$$\gamma(k) = \gamma(-k), \quad \rho(k) = \rho(-k).$$

Because ACF only describes the correlation between z_t and z_{t-k} and ignores the intermediate variable $z_{t-1}, \cdots, z_{t-k+1}$, we use PACF to describe this dependence on the intermediate variable. Let $\{z_t\}$ be a time series with $\mathbb{E}z_t = 0$, the PACF is defined as $\pi(1) = \mathrm{Corr}\left(z_1, z_2\right) = \rho(1)$ and

$$\pi(k) = \mathrm{Corr}\left(R_{1|2,\cdots,k}, \cdots, R_{k+1|2,\cdots,k}\right) \text{ for } k \geq 2,$$

where $R_{j|2,\cdots,k}$ is the residual from the linear regression of z_j on $(z_2, \cdots, z_k)$, namely

$$R_{1|2,\cdots,k} = z_j - \left(\alpha_{j2}z_2 + \cdots + \alpha_{jk}z_k\right),$$

and

$$\left(\alpha_{j2}, \cdots, \alpha_{jk}\right) = \underset{\beta_2,\cdots,\beta_k}{\arg\min}\, \mathbb{E}\{z_j - (\beta_2 z_2 + \cdots + \beta_k z_k)\}^2.$$

Given a time series observation $\{z_1, \cdots, z_T\}$, we can use the sample autocovariance function and sample autocorrelation function to estimate ACVF and ACF, which is defined as

$$\widehat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(z_t - \bar{z}_T)(z_{t+k} - \bar{z}_T), \quad \widehat{\rho}(k) = \widehat{\gamma}(k)/\widehat{\gamma}(0),$$

where $k = 0, 1, \cdots, T-1$ and $\bar{z}_T = \frac{1}{T}\sum_{t=1}^{T} z_t$.

B. IMAGING TIME SERIES

1) GRAMIAN ANGULAR FIELD

Given time series $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ with length n, time series data is scaled to between $[-1, 1]$ in the following way

$$\widetilde{x}_i = \frac{(x_i - \max(\boldsymbol{x})) + (x_i - \min(\boldsymbol{x}))}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})},$$

or scale to between $[0, 1]$ using the following formula

$$\widetilde{x}_i = \frac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}.$$

The shrinking time series data are converted into the form of polar coordinates, and the specific transformation formula is shown as follows [14]

$$\begin{cases} \phi = \arccos(\widetilde{x}_i), -1 \le \widetilde{x}_i \le 1, \\ r = \dfrac{t_i}{N}, \quad t_i \in \mathbb{N}, \end{cases}$$

where $\widetilde{x}_i \in \widetilde{\boldsymbol{x}} = \{\widetilde{x}_1, \cdots, \widetilde{x}_n\}$ and N is a constant factor to regularize the span of the polar coordinate system. The transformation into polar coordinates not only preserves the time series data changing with time and the related statistical properties but also presents the time series data in the form of graphs. The data of time series is converted to polar coordinates, and time can be easily determined by using the relation between angles. The definition of a GASF is as follows [14]

$$\mathrm{GASF} = \begin{pmatrix} \cos(\phi_1 + \phi_1) & \cdots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \cdots & \cos(\phi_2 + \phi_n) \\ \vdots & & \vdots \\ \cos(\phi_n + \phi_1) & \cdots & \cos(\phi_n + \phi_n) \end{pmatrix}$$
$$= \widetilde{\boldsymbol{x}}^\top \cdot \widetilde{\boldsymbol{x}} - \sqrt{\boldsymbol{1} - \widetilde{\boldsymbol{x}}^2}^\top \cdot \sqrt{\boldsymbol{1} - \widetilde{\boldsymbol{x}}^2}.$$

Similarly, the definition of the GADF is shown below

$$\mathrm{GADF} = \begin{pmatrix} \cos(\phi_1 - \phi_1) & \cdots & \cos(\phi_1 - \phi_n) \\ \cos(\phi_2 - \phi_1) & \cdots & \cos(\phi_2 - \phi_n) \\ \vdots & & \vdots \\ \cos(\phi_n - \phi_1) & \cdots & \cos(\phi_n - \phi_n) \end{pmatrix}$$
$$= \sqrt{\boldsymbol{1} - \widetilde{\boldsymbol{x}}^2}^\top \cdot \widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}^\top \cdot \sqrt{\boldsymbol{1} - \widetilde{\boldsymbol{x}}^2},$$

where $\boldsymbol{1} = [1, 1, \cdots, 1]$ is a row vector of all ones.

2) MARKOV TRANSITION FIELD

After given time series $\boldsymbol{x}$, the Q quantile of the family sequence is determined. Each observation x_i of the time series is allocated to the corresponding quantile interval $q_j (j \in [1, Q])$. Considering the first-order markov model along time, we can obtain the weighted adjacency matrix $\boldsymbol{W}$ of $Q \times Q$ and $\boldsymbol{W}_{i,j}$ is the frequency at which the point q_j is followed by q_i. When we normalize the matrix $\boldsymbol{W}$ by $\sum_j \boldsymbol{W}_{ij} = 1$, we obtain the Markov Transition matrix.

According to the definition of the Markov Transition matrix, it can be known that it is insensitive to the distribution of time series $\boldsymbol{x}$ and the moment t_i, which will lead to too much loss of information. In order to solve this problem, we define the following Markov Transition Field [14]

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{W}_{ij}|x_i \in q_i, x_1 \in q_j & \cdots & \boldsymbol{W}_{ij}|x_1 \in q_i, x_n \in q_j \\ \boldsymbol{W}_{ij}|x_2 \in q_i, x_1 \in q_j & \cdots & \boldsymbol{W}_{ij}|x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_{ij}|x_n \in q_i, x_1 \in q_j & \cdots & \boldsymbol{W}_{ij}|x_n \in q_i, x_n \in q_j \end{pmatrix}.$$

By assigning the time series data to Q quantile interval, we can obtain markov transition matrix $\boldsymbol{W}$. In MTF, $\boldsymbol{M}_{ij}$ is the probability of $q_i \rightarrow q_j$.

3) RECURRENCE PLOT

Time series data have obvious periodicity and unequal periodicity. Recursion of states in nonlinear systems or random processes is a typical scenario for generating time series. Recurrence plot [10] is a m dimensional space data in the 2-D tool. Its main idea is, at what point which tracks can return to the previous state, its mathematical expression for

$$\boldsymbol{R}_{i,j} = \theta(\epsilon - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|), \ \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{R}^m, \quad i, j = 1, 2, \cdots, K,$$

where K is the number of considered states, ϵ is the threshold distance. $\|\cdot\|$ is L_2 norm and θ is the Heaviside function. In the matrix $\boldsymbol{R}$, there is single point texture masonry, diagonal texture masonry, vertical line texture, and horizontal line texture. Moreover, the texture information has homogeneity, periodicity, drift, and fracture. Figure 3 shows the image obtained by encoding the time series.

C. FEATURE EXTRACTION

Based on ACF and PACF, we can determine the order p_i of the sequence SPEI$-i$ and construct the training set $\mathcal{D}_i = \left\{\left(\boldsymbol{x}_j^i, y_j^i\right)\right\}, i = 3, 6, 9, 12$ of the model, where $\boldsymbol{x}_j^i = \left(x_j^i, x_{j+1}^i, x_{j+p_i-1}^i\right) \in \mathcal{R}^{p_i}, y_j = x_{j+p_i}^i$ and $1 \le j \le n_i - p_i$. We can use the constructed data set $\mathcal{D}_i$ and various models, including SVR, RF for drought prediction, but the input of these models is the raw data, and these models cannot extract the features of the raw data well, especially on the data with special structure like time series.

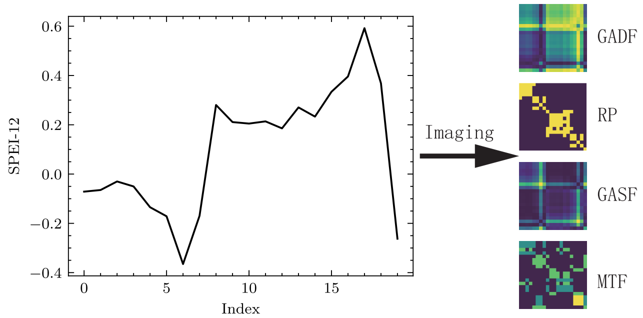

FIGURE 3. Examples of four time series imaging methods.

In this article, we use CV technology to extract features from $\mathcal{D}_i$, and use the extracted features for drought prediction. The most important step is to build a feature extraction network. The construction of the feature extraction network and the feature extraction based on transfer learning will be introduced respectively in III-C1 and III-C2.

1) CONSTRUCTION OF FEATURE EXTRACTION NETWORK

The construction of a feature extraction network based on CV requires a labeled image data set. We consider the images obtained by the same imaging technology as one class. If $\boldsymbol{x}_j^i$ in the data set $\mathcal{D}_i$ is directly encoded into an image, since $\boldsymbol{x}_j^i$ and $\boldsymbol{x}_{j+1}^i$ have a large amount of overlap, then the encoded image will also have a lot of redundancy, which is not conducive to learn a suitable feature extraction network. Therefore, considering dividing the SPEI-i data with length L_i, the images obtained by each imaging scheme are of one class. Generally speaking, the number of years of precipitation observation is limited, and L_i is generally not too small, so the number of each type of image obtained by SPEI-i imaging will be very limited. In order to learn a better feature extraction network, the article considers combining the images obtained by encoding the SPEI-i data of eight sites for training CNNs. Algorithm 1 gives the image data set a construction process.

Based on the data $\mathcal{A}_i$ generated in Algorithm 1, we can easily construct the following four image data sets

$$\mathcal{I}_i = \left\{ \left(\text{GASF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), 1\right), \left(\text{GADF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), 2\right), \left(\text{MTF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), 3\right), \left(\text{RP}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), 4\right) \right\},$$

where $1, 2, 3, 4$ represents different class, $i = 3, 6, 9, 12$ represents different time scales. Based on image data sets $\mathcal{I}_i$, we can train the feature extraction network. The feature extraction network selected in this article is ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, VGG-11, VGG-13, VGG-16, VGG-19 and the corresponding batch normalization (BN) version [17]. We should note that because the eight sites are all located in the Haihe River Basin, we assume that the data segmentation length of the SPEI of the same time scale is the same for different sites.

2) FEATURE EXTRACTION BASED ON TRANSFER LEARNING

Generally speaking, assume that the sample space of a machine learning task $\mathcal{T}$ is $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input space and output space, respectively, and the joint probability density function is $p(\boldsymbol{x}, y)$. A sample space and its distribution can be called a domain:

$$\mathcal{R} = (\mathcal{X}, \mathcal{Y}, p(\boldsymbol{x}, y)).$$

Algorithm 1 Image Data Set Construction Process

Require:
SPEI data set $\mathcal{D}_i^{\text{item}} = \{x_{i,k}^{\text{item}}\}$, $i = 3, 6, 9, 12$, $1 \leq k \leq n_{\text{item}}$,
item $\in$ {BeiJing, Datong, Fengning, Huailai, Yuxian, Zhangjiakou, Weixian, Tangshan};
Data segmentation length L_i;
Four imaging methods for sequence data: GASF, GADF, RP, MTF.
Ensure:
for item **in** {Datong, Weixian, $\cdots$,Tangshan}
  for $m = \{1, 2, \cdots, [\frac{n_{\text{item}}}{L}]\}$
    $\boldsymbol{a}_{i,m}^{\text{item}} = \left(x_{(m-1)\times L_i+1}^{\text{item}}, \cdots, x_{m\times L_i+1}^{\text{item}}\right)$;
    for coding **in** {GASF, GADF, RP, MTF}:
      coding $\left(\boldsymbol{a}_{i,m}^{\text{item}}\right)$
    end
  end
end
Output: $\mathcal{A}_i = \left\{\text{GASF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), \text{GADF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), \text{MTF}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right), \text{RP}\left(\boldsymbol{a}_{i,m}^{\text{item}}\right)\right\}$

Given two domains, if their input space $\mathcal{X}$, output space $\mathcal{Y}$, and probability distribution $p(\boldsymbol{x}, y)$ are not all the same, then the two domains are considered different [25]. Transfer learning refers to the process of knowledge transfer in two different domains, using the knowledge learned in the source domain $\mathcal{R}_S$ to help the learning task on the target domain $\mathcal{R}_T$. Transfer learning is divided into two types according to different transfer methods: inductive transfer learning and transductive transfer learning. Inductive transfer learning generally has the following two transfer methods [35]:

- Feature-based approach: The output of the pre-training model or the output of the intermediate hidden layer is directly added to the learning of the target task as a feature.
- Fine-tuning method: reuse part of the main components of the pre-training model on the target task, and fine-tune its parameters.

In this paper, feature-based inductive transfer learning is used for feature extraction, that is, the input of the prediction model is imaged, and the pre-trained feature extraction network is used to extract the features of the imaging data for prediction. We have also considered the impact of different levels of features on the results of drought prediction, and section V will give a detailed explanation. Feature extraction based on transfer learning is shown in Figure 4.

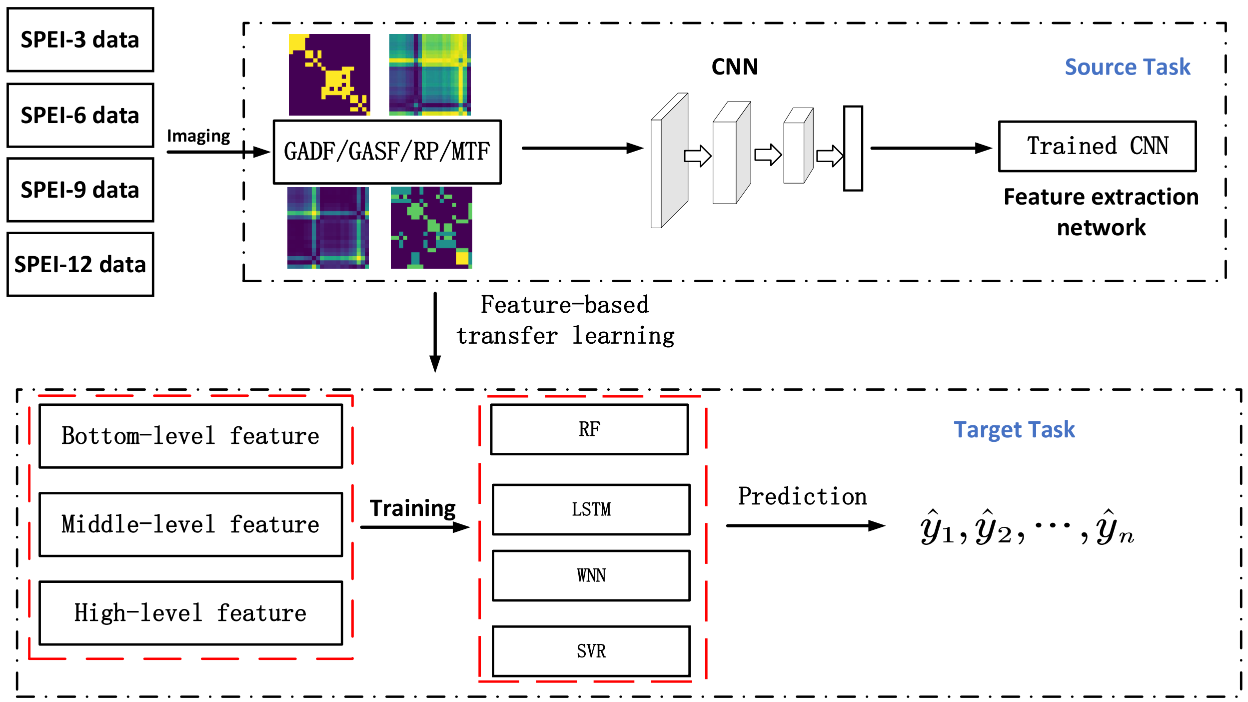

FIGURE 4. SPEI feature extraction based on transfer learning.

IV. FORECAST MODELS

The image features extracted by the feature extraction network will be used for drought prediction. In this article, we have selected four common models for drought prediction, including LSTM, RF, WNN, and SVR. Below, we will give a brief introduction to these methods.

A. RANDOM FOREST

The essence of RF [6] is an ensemble learning algorithm. The ensemble process used is bagging, and the base learner is a decision tree (DT). For many applications, RF and boosting have similar performance and are easy to train and tune. In ensemble learning, the main idea of bagging is to average many noisy and progressively unbiased models to reduce the variance of the model. Because the tree can capture the complex interaction terms in the variable, and its deviation gradually decreases as the depth of the tree increases and the tree model contains a lot of noise, the tree is an ideal base classifier for the ensemble. Algorithm 2 gives the running process of RF.

Algorithm 2 Random Forest for Regression

Require:

Data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$;

Number of decision trees B;

Ensure:

for $b = 1$ **to** B:

Get bootstrap sample $\mathcal{D}^*$ from $\mathcal{D}$;

Construct a random-forest tree T_b based on $\mathcal{D}_b$. Iteratively use the following steps for each

terminal node of the tree until the number of tree nodes reaches the minimum $K_{\min}$;

(I) Randomly select m variables from p variables;

(II) Choose the best split point from m variables;

(III) Split m variables into two child nodes.

Output decision tree sequence $\{T_b\}_1^B$.

end

Output: $\hat{f}_{\mathrm{rt}}^{B}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\boldsymbol{x})$

B. LONG SHORT-TERM MEMORY

LSTM [16] is a special RNN. LSTM neural network joins input gate $\boldsymbol{i}_t$, outputs $\boldsymbol{o}_t$, forget gate $\boldsymbol{f}_t$. The forget gate $\boldsymbol{f}_t$ determines how much the state cell $\boldsymbol{c}_{t-1}$ at time t affects the state cell $\boldsymbol{c}_t$. The input gate $\boldsymbol{i}_t$ determines how much input is retained at the time $t-1$, and the output gate $\boldsymbol{o}_t$ determines how much the state unit at time t will enter into output and then participate in the LSTM calculation at time t,

forget gate:

$$\boldsymbol{f}_t = \mathrm{sigmoid}\left(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f\right),$$

input gate:

$$\boldsymbol{i}_t = \mathrm{sigmoid}(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f),$$

memory unit:

$$\begin{aligned} \widehat{\boldsymbol{c}}_t &= \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c), \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \circ \widehat{\boldsymbol{c}}_t, \end{aligned}$$

output gate:

$$\boldsymbol{o}_t = \mathrm{sigmoid}(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o),$$

the final output:

$$\boldsymbol{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{c}_t),$$

where sigmoid and tanh are two activation functions with the form:

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}, \ \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

The above formulas are the forward calculation process of LSTM. Training the LSTM neural network is to find the weight matrices $\boldsymbol{W}_f, \boldsymbol{W}_i$, $\boldsymbol{W}_c$ and $\boldsymbol{W}_o$, each offset $\boldsymbol{b}_f$, $\boldsymbol{b}_i$, $\boldsymbol{b}_c$, $\boldsymbol{b}_o$, and other optimal parameters to make the training results as close to the real results as possible. The advantage of LSTM is that it can learn and remember longer sequences and does not rely on pre-specified window lag observations as input, so it can get excellent prediction results in time series analysis.

C. WAVELET NEURAL NETWORKS

WNN [3] is a new type of neural network that combines wavelet analysis (WA) with feedforward neural network (FNN). Of course, wavelet analysis is currently also applied to CNNs and graph neural networks (GNNs) [34]. In this study, we used a multi-dimensional WNN, and the output is a linear combination of wavelet elements. The output of the WNN is represented by the following formula

$$\hat{y} = w_{m+1}^{[2]} + \sum_{j=1}^{m} w_j^{[2]} \Psi_j(\boldsymbol{x}) + \sum_{i=1}^{p} w_i^{[0]} x_i,$$

where $\boldsymbol{x} = \{x_1, x_2, \cdots, x_p\}$ is the input vector and m is the number of hidden layers. $\Psi_j(\boldsymbol{x})$ is a multi-dimensional wavelet multiplied by p scalar wavelets, and its expression is as follows

$$\Psi_j(\boldsymbol{x}) = \prod_{i=1}^{p} \psi(z_{ij}),$$

and

$$z_{ij} = \frac{x_i - w_{(\xi)ij}^{[1]}}{w_{(\zeta)ij}^{[1]}}.$$

The above formula is equivalent to a transformation of the input x_i. We call $w_{(\xi)ij}^{[1]}$ and $w_{(\zeta)ij}^{[1]}$ the translation and dilation factors, respectively. The wavelet function we use in this article is Mexican Hat

$$\psi(z_{ij}) = (1 - z_{ij}^2) \exp\left\{-\frac{1}{2} z_{ij}^2\right\}.$$

The parameters that need to be learned in the WNN are $w_i^{[0]}, w_j^{[2]}, w_{(\xi)ij}^{[1]}, w_{(\zeta)ij}^{[1]}, w_{m+1}^{[2]}, i = 1, 2 \cdots, p, j = 1, 2, \cdots, m$. There are different algorithms for learning wavelet neural networks. In this article, we use the back propagation (BP) algorithm. In addition, the learning of WNN is sensitive to the initial value of the parameters. Then we use the following initialization method for translation and dilation factors [36]

$$w_{(\xi)ij}^{[1]} = 0.5(M_i + N_i), \ w_{(\zeta)ij}^{[1]} = 0.2(M_i - N_i),$$

where M_i and N_i are the maximum and minimum values of the column where the x_i is located. Since the initialization of the parameters $w_i^{[0]}$ and $w_j^{[2]}$ has little effect on the performance of the model, they are randomly selected from the uniform distribution from 0 to 1.

D. SUPPORT VECTOR REGRESSION

SVR [27] is an extension of support vector machines (SVM) in regression problems. SVM was originally proposed to solve classification problems. Similarly, SVR also has hard margins, soft margins (by introducing slack variables), and non-linearity (by using kernel techniques), and can transform the original problem into a dual problem for solution. In this article, we introduce and use the soft-margin SVR model. By introducing slack variables ξ_i and ξ_i^*, the soft margin SVR is equivalent to the following optimization problem

$$\text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$

$$\text{Subject to } \begin{cases} y_i - \boldsymbol{w}^\top \boldsymbol{x} - b \le \nu + \xi_i, \\ \boldsymbol{w}^\top \boldsymbol{x} + b - y_i \le \nu + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, \end{cases}$$

the constant $C > 0$ determines the trade-off between the flatness of $f = \boldsymbol{w}^\top \boldsymbol{x} + b$ and the amount up to which deviations larger than ν are tolerated. The above optimization problem can be easily transformed into its dual problem and solved.

V. RESULTS AND DISCUSSION

In this part, we will give the results based on the model RF, LSTM, WNN, and SVR to predict the four time scales SPEI: SPEI-3, SPEI-6, SPEI-9, and SPEI-12 for eight sites. In order to determine the order $p_i, i = 3, 6, 9, 12$ of the model, we first draw the ACF and PACF of the SPEI of the eight sites. Because of the limited space, we only show the images of ACF and PACF of SPEI-12.

According to Figure 5, we can clearly observe that the order of the SPEI-12 of the eight stations is around 10. SPEI-3, SPEI-6, and SPEI-9 also have similar results. Because the deep learning models are not so sensitive to the order, for the convenience of calculation, we set $p_3 = p_6 = p_9 = p_{12} = 10$.

A. FORECAST RESULTS BASED ON RAW DATA

In this part, we will give the fitting and prediction results of the four models mentioned in the article on the data sets $\mathcal{D}_i^{\text{item}}$, and evaluate the results based on the four model evaluation indicators.

$$\mathcal{D}_i^{\text{item}}, \ i = 3, 6, 9, 12, \quad \text{item} \in \{\text{Beijing, Datong}, \cdots, \text{Tangshan}\}.$$

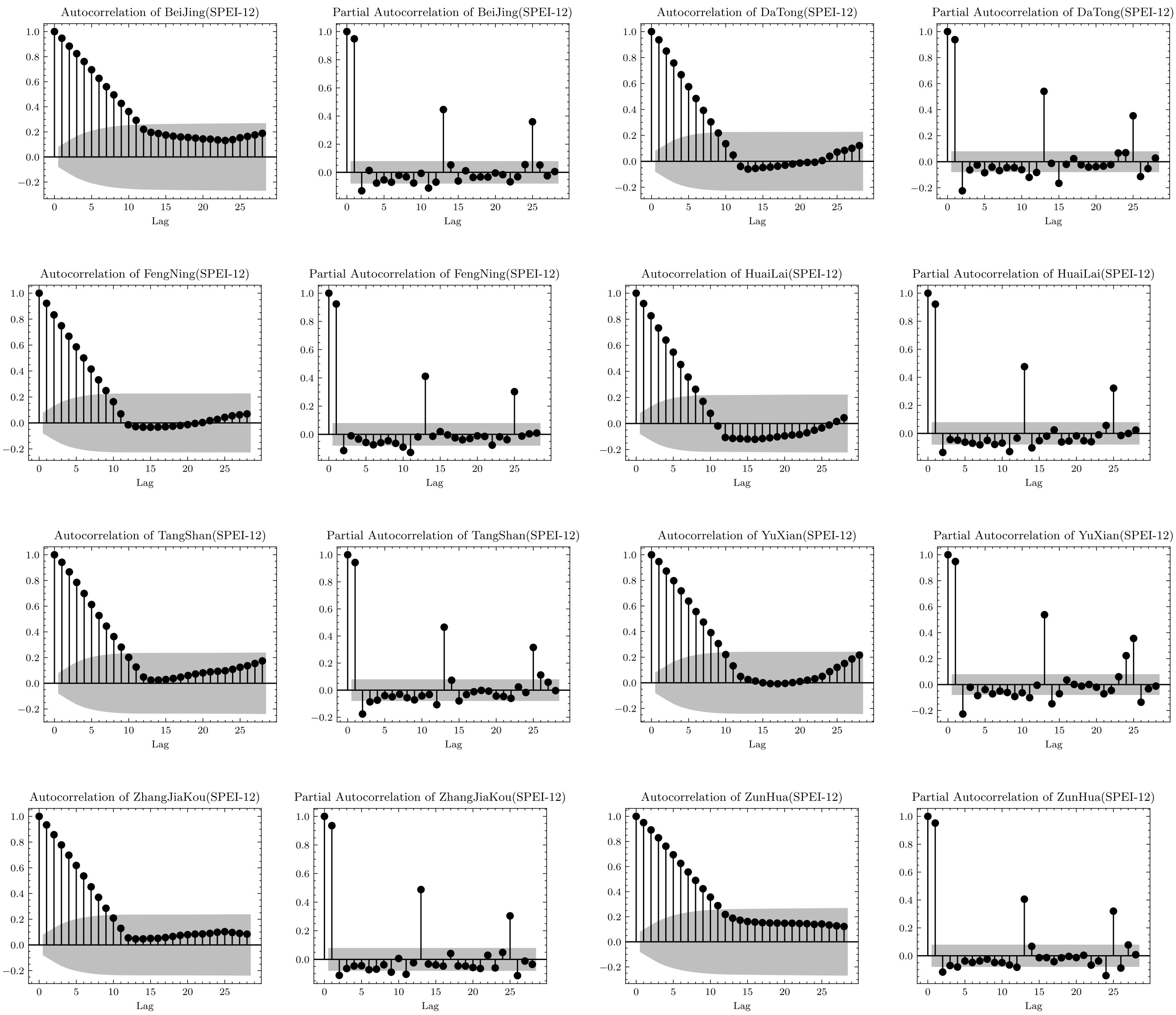

FIGURE 5. **ACF (left panel) and PACF (right panel) of the SPEI-12 sequence at eight stations.**

Due to limited space, we only show the prediction results on the test set. We should note that these models are not overfitted or underfitted in the test sets, and the results in the training set are similar. Table 3 to Table 6 show the experimental results in units of SPEI of different time scales.

From Table 3 to Table 6, we can get the following conclusions

1) In SPEI predictions at different time scales and sites, WNN and LSTM perform better than the other two models.
2) When the time scale is relatively small, such as SPEI-3 and SPEI-6, WNN performs better; while in SPEI-9 and SPEI-12, LSTM performs better. The main reason is that LSTM can capture long-term dependencies, while the WNN selected in this article has only one hidden layer, which can only capture short-term dependencies. The reason why the other two models perform worse than WNN and LSTM may be that they did not consider the dependence of each feature.
3) As the time scale increases, the performance of each model on SPEI gradually improves. For example, in the prediction of SPEI-3, the R^2 of each model does not exceed 0.5, but in the prediction of SPEI-12, the R^2 of each model can be over 0.9. Other model evaluation indicators, such as MAE, MAPE, SMAPE, and MSE, have similar patterns.
4) The SPEI prediction performance of each model at different time scales is quite different, which means that the four models selected in this article are greatly affected by the time scale of drought prediction.

We should note that SVR and RF are common in predictive models and are also the most excellent models. WNN and LSTM are also excellent models in sequence prediction. At different time scales, the prediction performance of these

TABLE 3. The prediction results of different models on SPEI-3.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	1.9765	2.3787	2.3224	2.2684	7.2247	12.1292	4.8646	1.6478
	SMAPE	1.4121	1.3475	1.4000	1.4132	1.4070	1.399	1.4273	1.3967
	MAE	**0.7518**	0.6948	0.7020	0.8010	0.6356	**0.6891**	0.7222	0.6861
	MSE	**0.8743**	0.6810	0.7667	0.9444	**0.7387**	0.8073	0.8002	**0.7076**
	R^2	0.4516	0.6210	0.6381	0.6201	**0.3584**	0.4839	0.6036	0.4569
LSTM	MAPE	3.0696	2.1095	**0.3630**	3.1725	12.1649	11.3911	4.8460	2.3103
	SMAPE	1.2544	1.1406	1.2767	**0.6491**	1.1997	1.1926	**1.0394**	1.1758
	MAE	1.1207	0.7811	0.7647	**0.2874**	0.7235	0.8318	**0.7205**	0.8944
	MSE	1.9037	0.9042	0.7776	1.2331	0.9602	1.0605	0.8724	1.4151
	R^2	0.2643	0.5016	0.6258	0.5600	0.2869	0.4350	0.5721	0.2059
WNN	MAPE	2.0856	**1.7870**	1.9017	3.1999	**4.1885**	6.8363	**3.4948**	**1.1536**
	SMAPE	**1.1019**	**1.0197**	**1.0303**	1.0986	**1.1023**	**1.1196**	1.1414	**1.1285**
	MAE	0.7949	**0.6605**	**0.6296**	0.8013	**0.6042**	0.7206	0.7341	0.6877
	MSE	1.0019	**0.6591**	**0.6261**	0.9262	0.7659	0.8968	**0.7965**	0.7677
	R^2	0.4420	**0.6459**	0.7085	0.6290	0.3577	0.4720	**0.6081**	0.4410
SVR	MAPE	**1.9655**	2.2615	2.4533	2.1187	7.2458	11.6990	4.3015	1.7521
	SMAPE	1.3897	1.3939	1.4130	**0.6491**	1.4297	1.4048	1.4547	1.4047
	MAE	0.7715	0.6865	0.6940	0.7739	0.7067	**0.7103**	0.7612	**0.6800**
	MSE	0.8874	0.7145	0.7144	**0.8584**	0.8727	0.8195	0.8747	0.7468
	R^2	**0.4534**	0.5973	**0.6595**	**0.6915**	0.2433	0.4676	0.5547	**0.4873**

TABLE 4. The prediction results of different models on SPEI-6.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	2.1867	2.3354	2.2479	1.7428	3.0748	**3.1196**	6.8581	1.5414
	SMAPE	1.2509	1.4010	1.3823	1.3750	1.4062	1.1534	1.4013	1.1930
	MAE	0.5284	0.5288	0.5762	**0.5817**	0.4595	0.4733	0.4877	**0.4396**
	MSE	0.4368	0.4477	0.5518	0.5229	0.3876	**0.3500**	0.3969	**0.3217**
	R^2	0.7500	0.8933	0.7310	0.7663	0.6254	0.6827	0.7718	**0.7989**
LSTM	MAPE	**1.1115**	1.7665	**1.1924**	1.2614	1.3972	5.1797	7.5878	**1.0077**
	SMAPE	**0.7671**	0.9626	1.0255	0.9549	0.9226	0.9997	**0.9569**	**0.7070**
	MAE	0.4868	0.6812	0.6044	0.6285	0.4489	0.5157	**0.4636**	0.4647
	MSE	0.3849	0.9502	0.5914	0.5737	0.4079	0.4148	**0.3886**	0.3383
	R^2	**0.7532**	0.4518	0.6877	0.7411	0.5920	0.6498	**0.7773**	0.7611
WNN	MAPE	1.1247	**1.2507**	1.3608	**1.1854**	**0.9486**	4.5355	6.8127	1.0317
	SMAPE	0.8552	**0.8953**	**0.9200**	**0.8263**	**0.4738**	**0.9551**	1.0282	0.7905
	MAE	**0.4752**	**0.4787**	0.5690	0.5878	**0.4025**	0.4838	0.5135	0.5196
	MSE	**0.3705**	**0.3849**	**0.5371**	0.5407	**0.3436**	0.3719	0.4518	0.3928
	R^2	0.7525	0.7804	0.7304	0.7651	**0.6830**	**0.6955**	0.7380	0.7518
SVR	MAPE	2.3223	2.4628	2.1708	1.7222	3.2459	3.2015	7.1180	1.5593
	SMAPE	1.2199	1.3347	1.4080	1.3619	1.4268	1.2272	1.4191	1.2668
	MAE	0.5456	0.6182	**0.5274**	0.5941	0.4488	0.5165	0.5443	0.4969
	MSE	0.4718	0.6226	0.5399	**0.5160**	0.3914	0.4149	0.5042	0.4038
	R^2	0.6755	0.6027	0.7321	**0.7707**	0.6439	0.6350	0.6973	0.7486

models is very different, which means that before using these models, we should first have a good representation of the data. This also reflects that the current forecasting model is sensitive to the time scale of SPEI. The following will introduce the use of CNNs to extract features from data and then use them for prediction.

B. FORECAST RESULTS BASED ON FEATURE EXTRACTION

In this part, the model prediction results based on the feature extraction of CNNs are given. The feature extraction networks trained in this article are VGG and ResNet, including VGG11, VGG11BN, VGG13, VGG13BN, VGG16, VGG16BN, VGG19, VGG19BN, where "BN" represents batch normalization; and ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. We believe that models with good classification performance on the imaging data set $\mathcal{I}_i, i = 3, 6, 9, 12$ indicate that the data set has better feature extraction capabilities, so we choose those models with better classification performance for feature extraction. Figure 6 shows the training loss of each model and the classification performance on the imaging datasets. We should note that when segmenting the SPEI index of each time scale, we have a hyperparameter $L_i, i = 3, 6, 9, 12$ that needs to

TABLE 5. The prediction results of different models on SPEI-9.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	2.7036	3.3464	1.9244	5.1496	2.6832	6.5996	3.0801	0.9677
	SMAPE	1.1121	1.3179	1.2936	1.3503	1.3162	0.9871	1.2215	0.9964
	MAE	0.3703	0.4445	0.4728	0.4591	0.3229	0.3500	0.4354	0.3713
	MSE	0.2542	0.3225	0.3452	0.3678	0.1882	0.2144	0.3463	0.2113
	R^2	0.7975	0.7344	0.8188	0.8134	0.7959	0.6938	0.7515	**0.8895**
LSTM	MAPE	**0.9934**	2.6691	0.9669	**2.6839**	1.1120	2.0935	2.3446	**0.4366**
	SMAPE	**0.6361**	0.7682	0.7595	**0.6338**	**0.7595**	0.8068	0.6418	**0.4315**
	MAE	0.3603	0.4246	0.4347	**0.3807**	**0.2551**	0.3287	0.3981	0.3198
	MSE	0.2141	0.3799	0.3270	0.2754	**0.1397**	0.1763	0.2662	**0.0857**
	R^2	**0.8394**	0.6988	0.8317	0.8648	**0.8405**	0.8064	0.8093	0.8731
WNN	MAPE	1.1822	**1.9755**	**0.6146**	3.6178	**0.9413**	**1.8716**	**1.6052**	0.4768
	SMAPE	0.6451	**0.6523**	**0.6586**	0.6839	0.8280	**0.7853**	**0.5691**	0.4797
	MAE	0.3687	**0.3484**	0.3719	0.4212	0.2923	**0.3003**	0.3600	0.3561
	MSE	0.2156	**0.2532**	0.2869	0.3277	0.1847	**0.1531**	0.2667	0.1985
	R^2	0.8218	**0.8152**	0.8464	0.8431	0.8063	**0.8289**	0.8121	0.8691
SVR	MAPE	2.9600	3.7139	2.0572	5.4426	2.7278	6.1639	2.8463	1.0462
	SMAPE	1.1281	1.2562	1.3261	1.3588	1.3899	1.1757	1.2126	0.9948
	MAE	**0.3395**	0.3726	**0.3519**	0.3860	0.2738	0.3016	0.3282	**0.3131**
	MSE	**0.2008**	0.2855	**0.2352**	**0.2646**	0.1682	0.1592	0.2011	0.1925
	R^2	0.8253	0.7643	**0.8740**	**0.8707**	0.8226	0.8033	0.8583	0.8650

TABLE 6. The prediction results of different models on SPEI-12.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	1.9917	1.9760	1.6503	2.2096	2.7859	12.0726	1.5793	1.2395
	SMAPE	1.0489	1.0741	1.2791	1.3618	1.3565	1.0891	1.1153	0.8847
	MAE	0.2430	0.2340	0.3400	0.3022	0.1973	0.2886	0.2619	0.1949
	MSE	0.1038	0.0981	0.2316	0.1837	0.0792	0.1373	0.1306	0.0822
	R^2	0.8869	0.8933	0.8752	0.8841	0.8975	0.8447	0.8732	0.9319
LSTM	MAPE	**0.5400**	0.5430	**0.3630**	0.8680	0.6724	**1.3951**	0.7029	**0.4709**
	SMAPE	**0.4452**	0.5028	**0.3874**	**0.6491**	**0.4738**	0.6629	**0.4916**	**0.2880**
	MAE	**0.2150**	0.2350	**0.2396**	0.2874	0.1652	**0.2221**	0.2501	0.2044
	MSE	0.0853	0.0955	0.1730	**0.1657**	**0.0518**	0.0990	0.1232	0.0857
	R^2	0.8960	**0.9056**	0.8990	0.8948	**0.9325**	0.8569	0.8769	0.9415
WNN	MAPE	0.5964	**0.4200**	0.4334	**0.8472**	**0.6361**	2.2527	**0.6360**	0.4806
	SMAPE	0.4717	**0.4326**	0.5566	0.6806	0.5459	**0.6197**	0.5086	0.3363
	MAE	0.2168	0.2236	0.2655	0.3275	0.1793	0.2251	0.2666	0.2394
	MSE	0.0828	0.0930	**0.1714**	0.1929	0.0553	**0.0891**	**0.1244**	0.0996
	R^2	0.9027	0.9032	**0.8994**	0.8759	0.9321	**0.8706**	**0.8854**	0.9241
SVR	MAPE	1.9606	2.0559	1.6541	2.2152	2.7925	12.0515	1.6153	1.2478
	SMAPE	1.0314	1.0462	1.2550	1.3467	1.3008	1.0616	1.0864	0.8968
	MAE	0.2081	**0.2082**	0.3308	0.3157	**0.1640**	0.2466	**0.2464**	**0.1852**
	MSE	**0.0790**	**0.0808**	0.2244	0.1836	0.0541	0.1085	0.1276	**0.0755**
	R^2	**0.9046**	0.9047	0.8703	0.8830	0.9286	0.8544	0.8724	**0.9394**

be determined. For the convenience of calculation, we set $L_3 = L_6 = L_9 = L_{12} = 20$. It can be seen from Figure 6 that for the imaging data of SPEI-3 and SPEI-6, VGG11BN should be selected as the feature extraction network, for the imaging data of SPEI-9, ResNet50 should be selected and for SPEI-12, ResNet18 should be selected.

In the following, we will show the prediction results based on the features extracted by the feature extraction network (FEN). Because we use CNNs as the FEN, before feature extraction on $\boldsymbol{x}_j^i$ in the data set $\mathcal{D}_i$, we first need to encode $\boldsymbol{x}_j^i$ into an image. The article mentions four time series imaging schemes. Due to limited computing power, we use GASF to encode $\boldsymbol{x}_j^i$ before feature extraction. After feature extraction, we can get the following data sets

$$\mathcal{D}_i^{\mathrm{FEN}_i,\mathrm{item}} = \left\{\left(\mathrm{FEN}_i\left(\boldsymbol{x}_j^{i,\mathrm{item}}\right), y_j^{i,\mathrm{item}}\right)\right\},$$

$1 \leq j \leq n_i - p_{i,\mathrm{item}}, \boldsymbol{x}_j^{i,\mathrm{item}} = \left(x_j^{i,\mathrm{item}}, \cdots, x_{j+p_{i,\mathrm{item}}-1}^{i,\mathrm{item}}\right) \in \mathcal{R}^{p_{i,\mathrm{item}}}$, where $\mathrm{FEN}_i, i = 3, 6, 9, 12$ represents the FEN corresponding to the SPEI index of different time scales, $\mathrm{FEN}_i(\boldsymbol{x}_j^i)$ represents the feature after feature extraction of the original data $\boldsymbol{x}_j^i$ using the FEN_i. Table 7 to Table 10 show the prediction performance of each model on data set $\mathcal{D}_i^{\mathrm{FEN}_i,\mathrm{item}}$.

From Table 7 to Table 10, we can get the following conclusions

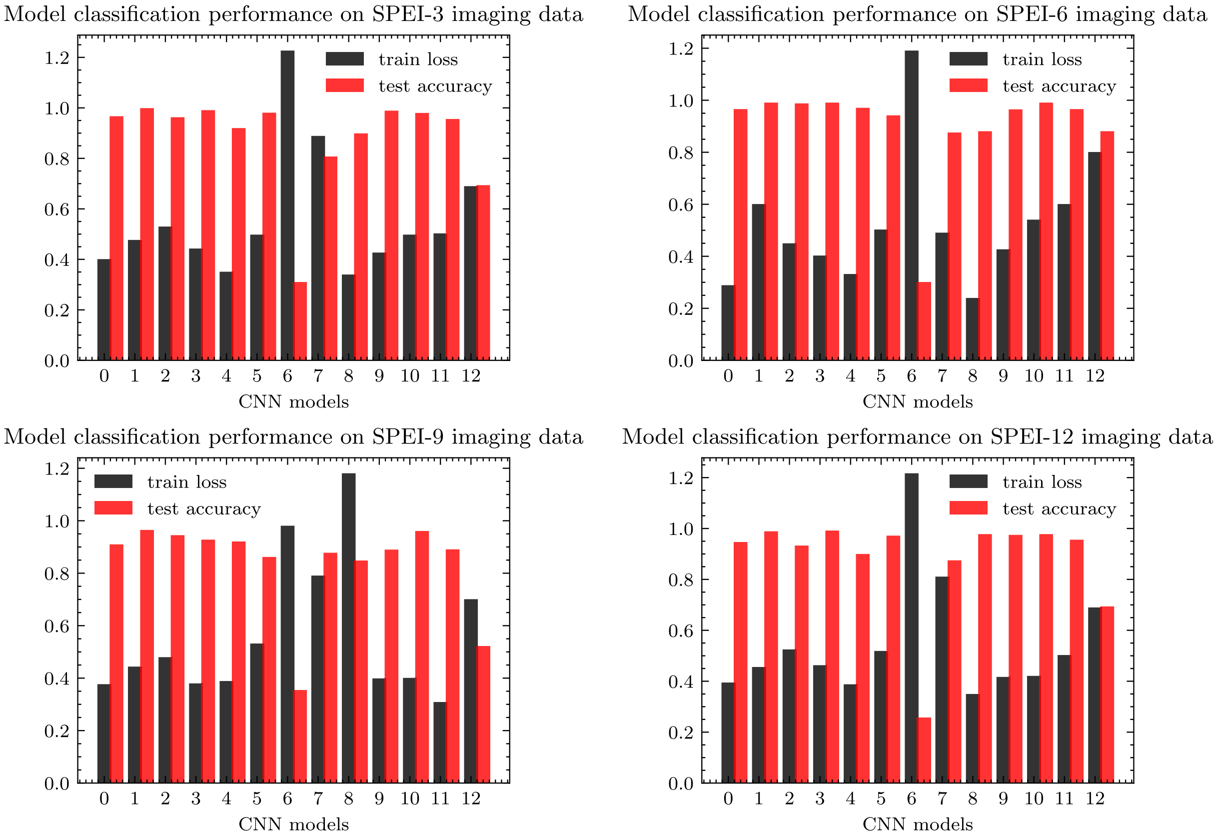

FIGURE 6. These four figures respectively show the classification performance of CNNs based on SPEI-3, SPEI-6, SPEI-9, and SPEI-12 imaging data. The horizontal axis 0 to 12 respectively represent VGG11, VGG11BN, VGG13, VGG13BN, VGG16, VGG16BN, VGG19, VGG19BN, ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152.

TABLE 7. The prediction results of different models on SPEI-3 with VGG11BN as FEN.

model	index \ site	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	**1.7789**	2.236	2.1831	2.1096	6.8635	11.4014	4.4268	1.4995
	SMAPE	1.285	1.2666	1.274	1.3284	1.2804	1.3291	1.3131	1.285
	MAE	0.7142	0.6601	0.6388	0.7369	0.6038	**0.634**	**0.65**	0.6381
	MSE	**0.7869**	0.647	0.7054	0.8688	**0.6648**	0.7427	0.7522	**0.6651**
	R^2	**0.5193**	0.7142	0.7402	0.6821	**0.3942**	0.5226	0.6941	0.5163
LSTM	MAPE	2.8547	1.9196	**0.334**	2.887	11.435	10.252	4.5068	2.1486
	SMAPE	1.154	1.0379	1.1618	**0.6166**	1.0917	1.1091	**0.9562**	1.1053
	MAE	1.0647	0.703	0.7188	**0.2587**	0.6801	0.7819	0.6701	0.805
	MSE	1.7133	0.859	0.7387	1.1345	0.893	0.9544	0.8288	1.3302
	R^2	0.3066	0.5367	0.7259	0.6048	0.3099	0.4698	0.6579	0.2409
WNN	MAPE	1.9188	**1.644**	1.7496	2.9439	**3.8953**	6.4261	**3.3201**	**1.0613**
	SMAPE	**1.0137**	**0.9177**	**0.9788**	0.9997	**1.0141**	**1.0188**	1.0387	**1.0157**
	MAE	0.7552	**0.6143**	**0.5729**	0.7212	**0.5679**	0.6846	0.6827	0.6464
	MSE	0.9318	**0.613**	**0.576**	0.8614	0.6893	0.834	**0.7487**	0.7216
	R^2	0.4818	**0.7169**	**0.8148**	0.7108	0.3899	0.505	**0.7115**	0.516
SVR	MAPE	1.8279	2.1484	2.257	1.9916	6.8111	10.6461	3.8714	1.6645
	SMAPE	1.2924	1.2545	1.3423	**0.6166**	1.2867	1.2784	1.3529	1.3064
	MAE	**0.7021**	0.6522	0.6385	0.7042	0.6431	**0.6464**	0.6927	0.6324
	MSE	0.843	0.6788	0.643	**0.8069**	0.7942	0.7621	0.8135	0.6945
	R^2	0.5123	0.6988	0.7189	**0.7606**	0.2798	0.5003	0.6046	**0.5604**

1) The performance of WNN and LSTM is better than that of RF and SVR in SPEI prediction at different time scales and sites.
2) At small time scales, such as 3 and 6, WNN performs better than LSTM, and at large time scales, such as 9 and 12, LSTM performs better. The main reason is that LSTM can capture long-term information.
3) As the time scale becomes larger, the prediction performance of each model on each site gradually improves. For example, based on the SVR to predict

TABLE 8. The prediction results of different models on SPEI-6 with VGG11BN as FEN.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	2.0118	2.1719	2.0681	1.6208	2.8288	**2.9636**	6.2409	1.4335
	SMAPE	1.1508	1.2749	1.2855	1.2375	1.3218	1.0496	1.2612	1.0976
	MAE	0.4756	0.4971	0.5243	0.5468	0.4181	0.4496	0.4389	0.4044
	MSE	0.415	**0.4164**	0.4966	0.4706	0.3605	**0.315**	0.3771	0.2927
	R^2	0.8475	0.8409	0.7895	0.8583	0.7067	0.751	0.8876	0.9187
LSTM	MAPE	**1.0226**	1.6782	**1.1328**	1.1857	1.2994	4.8689	7.2084	**0.917**
	SMAPE	**0.6904**	0.876	0.9332	0.9072	0.8765	0.9497	**0.9091**	**0.6434**
	MAE	0.4576	0.6267	0.5742	0.5908	0.4085	0.4641	**0.4172**	0.4182
	MSE	0.3657	0.8742	0.55	0.545	0.3712	0.3775	**0.3614**	**0.3214**
	R^2	0.821	0.4834	0.7565	0.8449	0.6394	0.6953	0.8395	0.8829
WNN	MAPE	1.0235	**1.1506**	1.2383	**1.0669**	**0.9012**	4.1727	6.2677	0.9595
	SMAPE	0.8039	**0.8505**	**0.8556**	**0.785**	**0.4501**	**0.8787**	0.9459	0.751
	MAE	**0.4467**	**0.4404**	0.5178	**0.5408**	**0.3622**	0.4403	0.4724	0.4676
	MSE	**0.3409**	0.3464	**0.4888**	0.5029	**0.3195**	0.3459	0.4247	**0.3732**
	R^2	**0.8804**	**0.874**	0.8034	**0.8722**	**0.765**	0.7442	0.8118	0.827
SVR	MAPE	2.1365	2.2411	2.0623	1.5672	2.9862	2.9454	6.4062	1.4501
	SMAPE	1.1467	1.2146	1.2813	1.2802	1.2841	1.1045	1.2914	1.1908
	MAE	0.502	0.5749	**0.501**	0.5644	0.4264	0.4855	0.5116	0.4571
	MSE	0.4482	0.5728	0.5129	**0.4696**	0.3601	0.3734	0.4538	0.3675
	R^2	0.7363	0.6871	0.8273	0.8324	0.7147	0.7302	**0.774**	**0.8235**

TABLE 9. The prediction results of different models on SPEI-9 with ResNet50 as FEN.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	2.3792	3.0118	1.7512	4.6861	2.388	5.9396	2.7413	0.8613
	SMAPE	1.012	1.1598	1.1513	1.1883	1.1714	0.8884	1.1116	0.9067
	MAE	0.3296	0.4045	0.4302	0.4132	0.2874	0.308	0.3919	0.3342
	MSE	0.2237	0.287	0.3141	0.331	0.1694	0.193	0.3151	0.1859
	R^2	0.7975	0.7638	0.8352	0.8297	0.7959	0.7007	0.7515	0.9073
LSTM	MAPE	**0.8742**	2.4289	0.8509	**2.3618**	0.9897	1.8423	2.0632	**0.3886**
	SMAPE	**0.5598**	0.6991	**0.676**	**0.5577**	**0.676**	0.71	0.584	**0.384**
	MAE	0.3279	0.3864	0.3869	**0.3426**	**0.227**	0.2958	0.3623	0.291
	MSE	0.1884	0.3419	0.2878	0.2424	**0.1271**	0.1551	0.2396	**0.078**
	R^2	**0.873**	0.7198	0.8567	0.8821	**0.8657**	0.7983	0.8012	**0.908**
WNN	MAPE	1.0758	**1.7977**	**0.5408**	3.2198	**0.8472**	**1.7032**	**1.4447**	0.4291
	SMAPE	**0.5741**	0.574	0.5862	0.6155	0.7535	**0.6989**	**0.5122**	0.4269
	MAE	0.3245	**0.3066**	0.3273	0.3791	0.266	**0.2733**	0.3168	0.3134
	MSE	0.194	0.2253	**0.2611**	0.2917	0.1644	**0.1347**	0.2427	0.1786
	R^2	0.8547	0.8315	0.8887	0.86	0.8386	**0.8538**	0.8283	0.8778
SVR	MAPE	2.664	3.3796	1.8309	4.8983	2.4005	5.5475	2.5901	0.9311
	SMAPE	1.0266	1.1055	1.167	1.2229	1.237	1.0346	1.0913	0.8854
	MAE	0.3056	0.3353	0.3132	0.3513	0.2492	0.2745	0.2954	**0.2818**
	MSE	**0.1827**	0.2541	**0.2117**	**0.2328**	0.1497	0.1401	0.183	0.1733
	R^2	0.8418	0.7796	**0.909**	**0.8881**	0.8391	0.7953	0.8926	0.891

at the Beijing site, the ranking of index R^2 can be obtained as 0.5123(SPEI-3) < 0.7363(SPEI-6) < 0.8418(SPEI-9) < 0.8956(SPEI-12).

By correspondingly comparing Table 3 to Table 6 with Table 7 to Table 10, we can get the following conclusions

1) Based on feature extraction, the prediction performance of each model on each time scale and site has been improved, and the performance improvement on SPEI3 and SPEI6 is the most obvious. For example, when the prediction model is RF and the data is SPEI-3 from Beijing site, the method based on feature extraction network has improved in each model evaluation index.
2) Compared with the prediction results of the original data, after feature extraction, the performance of each model on SPEI was balanced on each time scale. The performance of the models is not so different across sites and time scales. This means that feature-based extraction can significantly improve underperforming models.
3) In addition, feature extraction can reduce the model's sensitivity to time scales, that is, we have found a more general method for drought prediction.

TABLE 10. The prediction results of different models on SPEI-12 with ResNet18 as FEN.

model	site / index	Beijing	Datong	Fengning	Huailai	Tangshan	Yuxian	Zhangjiakou	Zunhua
RF	MAPE	1.7925	1.7586	1.5018	1.9886	2.5352	10.6239	1.3898	1.0908
	SMAPE	0.9335	0.9452	1.1512	1.2256	1.2344	0.9584	1.0038	0.8051
	MAE	0.2163	0.2129	0.2992	0.269	0.1736	0.254	0.2305	0.1715
	MSE	0.0934	0.0883	0.2038	0.1653	0.0697	0.1208	0.1175	0.0748
	R^2	0.8958	0.8844	0.919	0.9106	0.9154	0.8531	0.9081	0.9505
LSTM	MAPE	**0.4806**	0.4941	**0.3303**	0.7812	0.6119	**1.2277**	0.6256	**0.4238**
	SMAPE	**0.3918**	0.4525	**0.3409**	**0.5842**	**0.4169**	0.59	**0.4375**	**0.2563**
	MAE	**0.1892**	0.2092	**0.2108**	0.2529	0.147	0.1999	0.2251	0.1799
	MSE	0.0751	0.084	**0.1522**	**0.1508**	**0.0471**	0.0881	0.1109	0.0763
	R^2	0.896	0.9328	0.89	0.9306	0.9512	**0.8912**	0.912	0.9603
WNN	MAPE	0.5368	**0.3822**	0.3814	**0.7455**	**0.5598**	2.05	**0.5597**	0.4373
	SMAPE	0.4245	**0.3893**	0.4954	0.6057	0.4913	**0.5577**	0.4628	0.2959
	MAE	0.1908	0.1968	0.2336	0.2882	0.1596	**0.1981**	0.2373	0.2155
	MSE	0.0753	0.0846	0.1543	0.1698	0.0487	**0.0784**	**0.1107**	0.0896
	R^2	0.9478	0.9213	0.8994	**0.9197**	0.9694	0.8706	**0.9297**	0.9518
SVR	MAPE	1.7253	1.8503	1.4721	1.9937	2.5412	10.8459	1.4699	1.1105
	SMAPE	0.9179	0.9207	1.1295	1.1851	1.1707	0.9448	0.9886	0.8161
	MAE	0.1894	0.1832	0.2911	0.281	**0.1443**	0.217	**0.2168**	**0.1648**
	MSE	**0.0703**	0.0727	0.1997	0.1616	0.0492	0.0976	0.1136	**0.0664**
	R^2	0.8956	0.9137	0.9051	0.9095	**0.975**	0.8544	0.916	**0.977**

VI. CONCLUSION

The FENs used in this article are ResNet and VGG and their variants, but it is uncertain whether these two CNNs are the optimal feature extraction networks. For different data sets, we can adopt some other feature extraction networks, such as LeNet [18], GoogleNet [29], etc. Secondly, we need to pay attention to that because of the limited computing power, we set the hyperparameter segmentation length $L_3 = L_6 = L_9 = L_{12} = 20$ for each site and time scale SPEI. This may be unreasonable for personal data sets and should be set reasonably. We should also note that we set $p_i^{\text{item}} = 10$. In fact, it is clear from the ACF and PACF images that the order of the SPEI is not 10 for every time scale at every site, which should also be selected based on individual data. In addition, there are many other time series imaging technologies, such as Grey Scale Encoding [32], Spectrogram [5], etc. These imaging technologies can be used to expand the image data sets $\mathcal{I}_i, i = 3, 6, 9, 12$ in order to learn more generalized FENs. In general, this article only proposes a general method of precipitation prediction, whether it is data preprocessing, prediction model selection, or feature extraction network selection, it can be freely combined and innovated.

ACKNOWLEDGMENT

The authors would like to thank the editor, the AE, and the reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this article.

REFERENCES

[1] A. Abbasi, K. Khalili, J. Behmanesh, and A. Shirzad, "Drought monitoring and prediction using SPEI index and gene expression programming model in the west of Urmia Lake," *Theor. Appl. Climatol.*, vol. 138, nos. 1–2, pp. 553–567, Oct. 2019.

[2] A. Abebe and G. Foerch, "Stochastic simulation of the severity of hydrological drought," *Water Environ. J.*, vol. 22, no. 1, pp. 2–10, Mar. 2008.

[3] A. K. Alexandridis and A. D. Zapranis, "Wavelet neural networks: A practical guide," *Neural Netw.*, vol. 42, pp. 1–27, Jun. 2013.

[4] A. Belayneh, J. Adamowski, B. Khalil, and B. Ozga-Zielinski, "Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models," *J. Hydrol.*, vol. 508, pp. 418–429, Jan. 2014.

[5] B. Boashash, *Time-Frequency Signal Analysis and Processing*, 2nd ed. Oxford, U.K.: Academic, 2016.

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[7] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, "Large-scale malware classification using random projections and neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3422–3426.

[8] A. Dai, "Drought under global warming: A review," *Wiley Interdiscipl. Rev., Climate Change*, vol. 2, no. 1, pp. 45–65, Jan./Feb. 2011.

[9] Y. Deldjoo, N. T. Di, and F. A. Merra, "Adversarial machine learning in recommender systems (AML-RecSys)," in *Proc. WSDM*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 869–872.

[10] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.*, vol. 4, no. 9, pp. 973–977, 1987.

[11] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York, NY, USA: Springer, 2003.

[12] K. F. Fung, Y. F. Huang, C. H. Koo, and Y. W. Soh, "Drought forecasting: A review of modelling approaches 2007–2017," *J. Water Climate Change*, vol. 11, no. 3, pp. 771–799, Sep. 2020.

[13] P. Han, P. X. Wang, S. Y. Zhang, and D. H. Zhu, "Drought forecasting based on the remote sensing data using ARIMA models," *Math. Comput. Model.*, vol. 51, nos. 11–12, pp. 1398–1403, 2010.

[14] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," 2017, *arXiv:1710.00886*. [Online]. Available: http://arxiv.org/abs/1710.00886

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 448–456.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[19] J. Li, S. Zhou, and R. Hu, "Hydrological drought class transition using SPI and SRI time series by loglinear regression," *Water Resour. Manage.*, vol. 30, no. 2, pp. 669–684, Jan. 2016.
[20] M. Svoboda, M. Hayes, and D. A. Wood, "Standardized precipitation index user guide," World Meteorol. Org., Geneva, Switzerland, Tech. Rep. 1090, 2012.
[21] L. Meng, T. Ford, and Y. Guo, "Logistic regression analysis of drought persistence in East China," *Int. J. Climatol.*, vol. 37, no. 3, pp. 1444–1455, Mar. 2017.
[22] A. K. Mishra and V. P. Singh, "A review of drought concepts," *J. Hydrol.*, vol. 391, nos. 1–2, pp. 202–216, Sep. 2010.
[23] A. K. Mishra, V. R. Desai, and V. P. Singh, "Drought forecasting using a hybrid stochastic and neural network model," *J. Hydrol. Eng.*, vol. 12, no. 6, pp. 626–638, Nov. 2007.
[24] S. Park, J. Im, E. Jang, and J. Rhee, "Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions," *Agricult. Forest Meteorol.*, vol. 216, pp. 157–169, Jan. 2016.
[25] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA: Association for Computing Machinery, 2007, pp. 759–766.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–14.
[27] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[28] J. H. Stagge, I. Kohn, L. M. Tallaksen, and K. Stahl, "Modeling drought impact occurrence based on meteorological drought indices in Europe," *J. Hydrol.*, vol. 530, pp. 37–50, Nov. 2015.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[30] R. Taormina, K.-W. Chau, and B. Sivakumar, "Neural network river forecasting through baseflow separation and binary-coded swarm optimization," *J. Hydrol.*, vol. 529, pp. 1788–1797, Oct. 2015.
[31] H. Tatli, "Downscaling standardized precipitation index via model output statistics," *Atmósfera*, vol. 28, no. 2, pp. 83–98, Apr. 2015.
[32] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[33] B. Xu, R. Cai, Z. Zhang, X. Yang, Z. Hao, Z. Li, and Z. Liang, "NADAQ: Natural language database querying based on deep learning," *IEEE Access*, vol. 7, pp. 35012–35017, 2019.
[34] B. Xu, H. Shen, Q. Cao, Y. Qiu, and X. Cheng, "Graph wavelet neural network," in *Proc. IEEE Conf. Int. Conf. Learn. Represent. (ICIR)*, Jan. 2019, pp. 1–13.
[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 3320–3328.
[36] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Netw.*, vol. 3, no. 6, pp. 889–898, Nov. 1992.
[37] Y. Zhang, W. Li, Q. Chen, X. Pu, and L. Xiang, "Multi-models for SPI drought forecasting in the north of Haihe River Basin, China," *Stochastic Environ. Res. Risk Assessment*, vol. 31, pp. 2471–2481, Jul. 2017.
[38] Y. Zhu, S. Drake, H. Lü, and J. Xia, "Analysis of temporal and spatial differences in eco-environmental carrying capacity related to water in the Haihe River Basins, China," *Water Resour. Manage.*, vol. 24, no. 6, pp. 1089–1105, Apr. 2010.
[39] M. Özger, A. K. Mishra, and V. P. Singh, "Estimating palmer drought severity index using a wavelet fuzzy logic model based on meteorological variables," *Int. J. Climatol.*, vol. 31, no. 13, pp. 2021–2032, Nov. 2011.

WAN TIAN received the B.Sc. degree in applied chemistry from the China University of Petroleum (East China). He is currently pursuing the master's degree in applied statistics with Capital Normal University. His research interests include machine learning and computer vision.

JIUJING WU received the B.Sc. degree in mathematics and applied mathematics from Capital Normal University, where she is currently pursuing the master's degree in applied statistics. Her research interests include machine learning, non-linear correlation measures, and variable selection.

HENGJIAN CUI received the B.S. and M.S. degrees in mathematics from Beijing Normal University (BNU), China, in July 1983 and July 1986, respectively, and the Ph.D. degree in statistics from the Institute of System Sciences, Chinese Academy of Sciences, in July 1993. He is currently a Professor of statistics with the School of Mathematical Sciences, Capital Normal University, Beijing, China. He was a Visiting Scholar with the Department of Statistics, University of Illinois at Urbana–Champaign (UIUC), and the Department of Mathematics, Arizona State University (ASU), from July 2000 to August 2001. His research interests include parametric estimation theory in statistical linear model, growth curve model, measurement error (errors-in-variables) model, projection pursuit approach, hypothesis testing, multi-variate data analysis, non-parametric, semi-parametric and robust statistics, empirical likelihood method, quality control, statistical computing, and numerical analysis. His ongoing research interests include model checking and the theory of statistical depth functions and its applications to multivariate exploratory data analysis, pattern recognition, quality control and system reliability, and biomedical sciences.

TAO HU is currently a Professor of statistics with the School of Mathematical Sciences, Capital Normal University, Beijing, China. His research interests include mathematical statistics and applied statistics.