IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2019, accepted February 21, 2019, date of publication February 26, 2019, date of current version March 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901842

Stock Market Trend Prediction Using High-Order Information of Time Series

MIN WEN[1], PING LI[1], LINGFEI ZHANG[2], AND YAN CHEN[1]

[1]School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
[2]Corporate IT Department, Nomura Securities Corporation, Tokyo 103-8011, Japan

Corresponding author: Ping Li (dping.li@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873218 and Grant 61503312, and in part by the Applied Basic Research Foundation of the Sichuan Provincial Science and Technology Department under Grant 18YYJC1147.

ABSTRACT Given a financial time series such as *S&P* 500, or any historical data in stock markets, how can we obtain useful information from recent transaction data to predict the ups and downs at the next moment? Recent work on this issue shows initial evidence that machine learning techniques are capable of identifying (non-linear) dependency in the stock market price sequences. However, due to the high volatility and non-stationary nature of the stock market, forecasting the trend of a financial time series remains a big challenge. In this paper, we introduced a new method to simplify noisy-filled financial temporal series via sequence reconstruction by leveraging motifs (frequent patterns), and then utilize a convolutional neural network to capture spatial structure of time series. The experimental results show the efficiency of our proposed method in feature learning and outperformance with 4%–7% accuracy improvement compared with the traditional signal process methods and frequency trading patterns modeling approach with deep learning in stock trend prediction.

INDEX TERMS Trend prediction, convolutional neural network, financial time series, motif extraction.

I. INTRODUCTION

In financial stock market, time series is a sequence of the price of a given share. It is of particular importance for the investment to predict the trend of financial time series. Compared to other time series, such financial time series possesses some special characteristics owing to the microstructure of the financial market. One basic feature is the high frequency of individual values, leading to the intensification of the influence of non-systematic factors to the dynamics of those time series. As a result, financial time series exhibit high volatility and non-stationarity. Due to its inherent uncertainty, the task of forecasting financial time series has been proved to be challenging, where traditional statistical models [20]–[22], machine learning methods [7] and artificial neural networks(ANN) [2]–[4] have been widely investigated. Most of the existing models are highly focusing on precise prediction of future prices. On the other hand, there are some of research being conducted regarding the forecasting of the direction of the prices, i.e., the up and down trends of the time series. Trend forecasting is more concerned with future volatility trends than accurate price forecasts. It provides investors with a decision message from another perspective.

Recently, some new techniques have been developed to tackle this problem. For instance, Ding *et al.* [25] and their series of studies [26], [27] combined Deep Learning(DL) technology and Natural Language Processing(NLP) approach. From the perspective of external information in financial markets, they use financial news to study the impact of stock market volatility. Specifically, they use the named entity recognition technology to process financial news title and extract features with Neural Tensor Network(NTN) from news entities. By using a multi-layer NTN model, they predict S&P 500 index and individual stock price trend. Different from traditional analysis means and research direction, their method can effectively capture market trends based on financial news. However, their work relies heavily on the quality and real-time nature of news reporting, which limits the feasibility and accuracy of the model.

Since there are too many factors such as public opinions, general economic conditions, or political events, that have direct or indirect impacts on the evolution of financial time series, extracting these features is tedious and costly. Can we acquire useful information by purely analyzing the original

The associate editor coordinating the review of this manuscript and approving it for publication was Zhong-Ke Gao.

time series? To this end, in the present work we propose a simply but efficient method to capture the latent patterns in the financial time series and consequently make predictions on the up/down trends. More precisely, we focus our attention on a way of sequence reconstruction by leveraging higher-order characteristic named as motifs (frequent and meaningful patterns) existing in financial series. We propose a three-step method for trend prediction: firstly, discovering the dynamic evolutionary motif, followed with reconstructing the sequence by stitching the motif sequence according to the order of discovery, finally applying convolutional neural network (CNN) on the reconstructed sequence's feature and carrying out the trend prediction.

In summary, our paper makes two contributions to this literature:

1) We propose a new way on financial sequence reconstruction: most of previous methods focus on frequency decomposition of time series signals or phase space reconstruction such as wavelet decomposition or Empirical Mode Decomposition. However, this kind of processing method was affected by local characteristics of data and lack of consideration of dependencies between time segments which result in inaccurate performance of prediction model. In order to reduce this effect, we simplify redundant and noisy-filled financial temporal data by leveraging motif information existing in financial series.
2) We present a new trend prediction scheme on financial time series: Different from the existing sequence prediction models, we utilize CNN combined with two full-connected layers to capture the spatial correlations between historical and current trends and achieve an improvement in the performance.

The rest of this paper is arranged as follows: In Section 2, we give a brief review of related works, some of them will be compared with our framework. Then we describe the scheme of the proposed approach and parameter settings related to the neural network model in detail in Section 3. After that, we conduct experiments and compare the method with baselines to evaluate the performance of our model in Section 4. Finally, we close with a discussion about this paper and future work in Section 5.

II. RELATED WORK

Many efforts to financial time series analysis problem have focused on price or trend prediction, and they roughly include traditional time series analysis methods, machine learning methods, Deep learning methods and their combination.

Pioneering work in the problem mainly used traditional time series analysis methods and the means of signal processing. Box *et al.* [14] proposed classical time series analysis method for time series on both stationary and non-stationary condition. One of the most prominent univariate time series models is the autoregressive integrated moving average (ARIMA) model. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box-Jenkins methodology in the model selection procedure. However, the effect of this model depends on the smoothing of non-stationary time series data, especially on nonlinear financial time series.

On the other hand, some of signal processing technology have been applied in non-stationary time series analysis. Reference [11] designed a robust finite-horizon filters which could provide a better transient filtering performance if the noise inputs are nonstationary. References [20]–[22] Combined Stationary Wavelet Transform with some prediction methodologies provide a new way to deal with financial time series and it has been used for stock price prediction. Recent research [39] shows that bridging wavelet multi-resolution and multivariate complex network [40] provides a novel methodology for analyzing multivariate nonlinear time series which widely exist in nonlinear science and engineering. In addition, the missing data problem in fragmented time series is a common challenge in trend analysis. Most classical methods replacing missing data with constants or straight line values will violate those assumptions about the noise statistics. Reference [10] use hypothesis-testing-based adaptive spline filtering (HASF) algorithm to discover the trends of fragmented time series for mHealth Apps, which can accommodate non-uniform sampling and is therefore inherently robust to missing data.

Non-stationary time series show varying degrees of volatility in some periods, and these volatility depend on the past of the sequence. Some other methods like markov chain and recurrent neural network provide us different approach to model the temporal dependence of non-stationary time series, for instance in financial field. As a special financial time series application, stock market prediction is a classic problem which has been studied extensively using various machine learning tools and techniques. Tay and Cao [15] and Kim [16] studied various parameters(the upper bound and the kernel parameter) of SVM to select optimal values for the best prediction performance. In addition to this, the work of Hassan and Nath [17] and Gupta and Dhingra [18] shows that Hidden Markov Model(HMM) is widely used in financial time series price forecasting and is better than ANN and ARIMA method in price predicting accuracy. However, the hyper-parameter settings (such as the number of hidden state) of HMM is easily influenced by people and there is no authoritative principle to guide the selection of parameters.

Recent advances in computing power and the development of Deep learning [36] enable more novel models to be applied in price prediction. According to the timing dependence of non-stationary time series, sequence model, such as recurrent neural network(RNN) with its variant [35], are frequently used in processing time series data. For instance, Xiong *et al.* [24] devised a LSTM neural network to model S&P 500 volatility incorporating Google domestic trends. Similarly, Fischer and Krauss [37] applied LSTM networks to a large-scale financial market trend and price prediction task. However, both of them are leverage LSTM model by

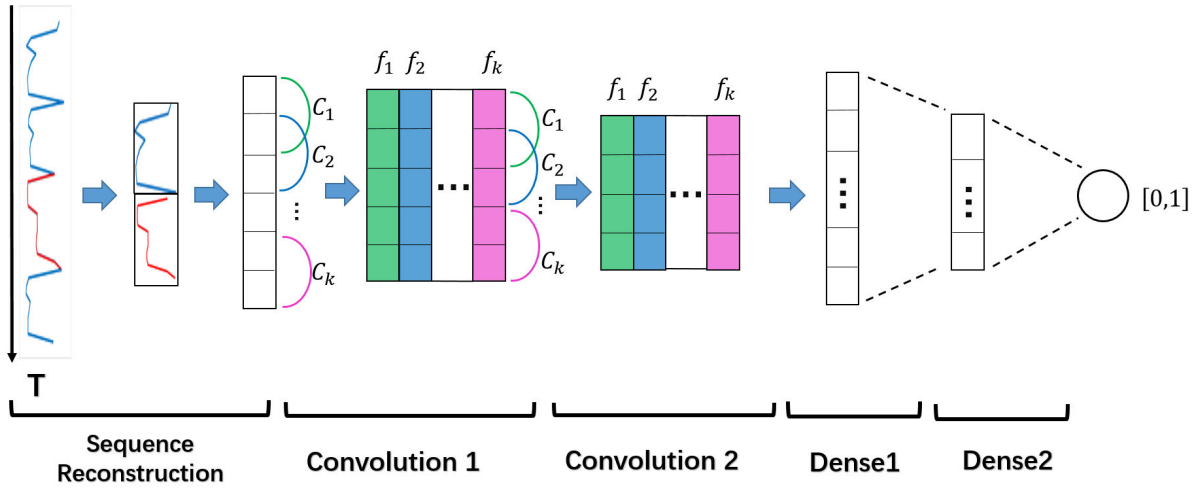

FIGURE 1. Architecture of the financial trend prediction model.

accumulate or forget information in a long period of time to modeling the temporal relationship of sequence data.

There are two major challenges associated with forecasting financial time series: (1) nonstationary (i.e., the statistical properties of the time series change with time); (2) multi-scaling (i.e., the statistical properties of the time series change with time-horizons). It is not convenient to model these volatility effect directly, hence some methods combined neural network with Empirical Mode Decomposition(EMD) [6]–[8] and detrending [1] has been used to solve this problem indirectly. The general idea behind the use of EMD for forecasting purposes is therefore that by dividing a signal into intrinsic mode functions(IMF) components, the residual component can reduce the complexity of the time series, separating trends and oscillations at different scales, improving in this way the forecasting accuracy at specific time-horizons. In addition to this, pattern recognition of temporal also has lunch a new research direction. There are some work [5], [9], [28]–[33] that exploring the extraction of important characteristics or patterns from financial time series received a lot attention. Zhang *et al.* [33] proposed a novel State Frequency Memory(SFM) recurrent network to capture the multi-frequency trading patterns from past market data to make long and short term predictions over time. Specifically, SFM was used to decompose the hidden states of memory cells into multiple frequency components, each of which models a particular frequency of latent trading pattern underlying the fluctuation of stock price. However, most of these works use neural network to extract frequent pattern form Multivariate time series and combine with regression or classification model to do prediction but lack of consideration of frequent patterns with data.

Overall, current research is mainly concerned with modeling the correlation from time series dependence(i.e., LSTM, HMM), signal decomposition of time series(i.e., ARIMA, EMD, wavelet transform) and time series pattern recognition(i.e., SFM). Differing from previous work, we proposed a unique denoise method of financial temporal series by sequence reconstruction technology using frequent pattern recognize and adopted CNN instead RNN for feature extraction combining with full-connected neural network for stock price trend prediction. It is a simple but useful method compared to previous work and the details will be described in the next part.

III. THE PROPOSED METHOD

In this work, we proposed an algorithm combining motif-based sequence reconstruction with CNN for stock time series trend prediction. Our basic idea is that the trends of a time series can be reflected in the patterns of diverse motifs. Therefore, a better representation of financial time series for trend prediction should involve high-order structural features, e.g., motifs. To this end, we design a two-stage predictive model by representing time series at a high level, as shown in Figure 1. Specifically, given a financial time series, we extract representative motifs to summarize the financial time series rather than process on raw data. Next, we use CNN model to extract high-order abstract features from frequent patterns by leveraging the enhanced expressiveness of the reconstructed sequence. The overall process can be divided into the following steps:

1) For every single time series, acquiring a set of motifs using the vocabulary-based model and Modified Dynamic Time Warping(MDTW) distance metrics presented in [31].
2) Reconstructing original sequence by stitching all frequent motifs according to the order in which these frequent items appear.
3) Feeding them into convolutional neural networks for feature selection, after that, a fully connected neural network with one output and a Dropout layer is used for trend prediction.

Algorithm 1 Vocabulary-Based Motif Extraction

Input: Time series X
Output: Vocabulary V_1,...,V_k

$k = 0$
$i = 1$ //current position
while $i < m$ **do**
  ▷ Find best vocabulary-prefix match, see [31] Section 6.2
  $j^*, s^* = \arg\min_{j,s} \frac{MDTW(X_{i+1:i+s}, V_j)}{s}$
  if $MDTW(X_{i+1:i+s}, V_j) < C_F \cdot s^*$ or $k = k_{max}$ **then**
    ▷ VocabuMerge, see [31] Section 6.3
    $V_{j^*} = VocabuMerge(X_{i+1:i+s^*, V_{j^*}})$
    Append j^* to Z
  else
    ▷ Create new vocabulary term, see [31] Section 6.4
    $s^* = CreateNewVocabu(X, i, (V_1, ..., V_k))$
    $V_{k+1} = X_{i+1:i+s^*}$
    Append $k + 1$ to Z
    $k = k + 1$
  end if
  Append $[i + 1 : i + s^*]$ to S
  $i = i + s^*$
end while
Return V_1,...,V_k

TABLE 1. Symbol description.

Symbol	Interpretation
X	Raw data of input time series
V_j	jth motif
l	window size of series
m	Length of X
k	Number of motifs
k_{max}	Maximum motif size
s_{min}, s_{max}	Minimum and maximum width of segment
$Z(i)$	motif assignment for ith motif
C_F	Average standard deviation of input data
MDTW	Modified DTW

A. MOTIF EXTRACTION

First we list the parameters in Table 1.

In the following, we describe how to perform motif extraction. Given a series X, a motif set V, motif extraction aims at segmenting time series into subsequence set S and assigning motif labels to each element of S. Other parameters involved are listed in Table 1. As figure 2 shows, each motif represents a frequently occurring subsequence corresponding to specific pattern. Z records the motif to which each segment is assigned. S includes the location of the data segmentation.

1) Segmentation S:this describes how series X is splitted into continue segments of time. Using $[a_i, b_i]$ as segment term, where $a_1 = 1$, $b_1 = l$, and $b_i + 1 = a_{i+1}$ for $i = 1, ..., m - 1$.
2) Assignment variables Z: this describes which vocabulary term is used to encode each segment. For each i, the assignment variable $Z(i)$ means that the $Z(i)$th motif (i.e. $V_{Z(i)}$) is used to encode segment i.

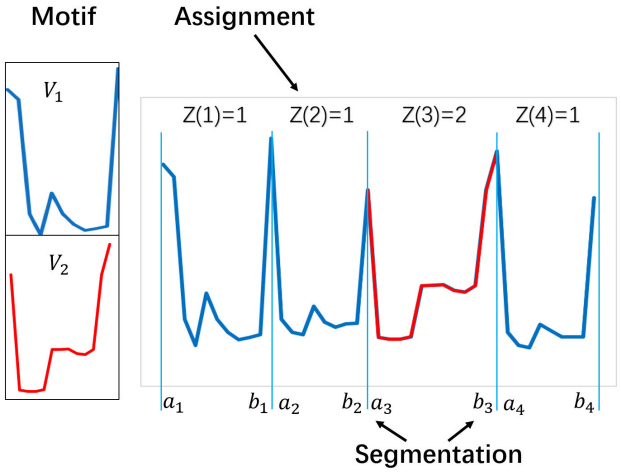

FIGURE 2. An example of time series segmentation.

3) MDTW distance: we use MDTW distance as a measure of the similarity of two sequences. Given two sequences $x_1 = X_{i_1,j_1}$ and $x_2 = X_{i_2,j_2}$, MDTW(x_1, x_2) modifies regular DTW by adding a penalty of $\log n_{x_1}$ for each expansion to x_1 and $\log n_{x_2}$ for each expansion to x_2, where n_{x_1} and n_{x_2} are the lengths of x_1 and x_2.

First, we generate the initial motif whose length is in range (s_{min}, s_{max}) which is selected according to MDL [38] method. Then, starting at the beginning of the rest of time series, our algorithm repeatedly finds the closest match between any existing motif where distance metrics is average MDTW and the length of scan is limit to (s_{min}, s_{max}). If a sufficiently good match is found, it assigns this segment to this motif. Otherwise, it creates a new motif. Algorithm 1 describes the whole process of motif extraction.

B. SEQUENCE RECONSTRUCTION

Financial time series is generally recognized as chaotic time series and traditional series reconstruction is built in the phase space [19]. In contrast, we reassembly the time series using the motif set V extracted by algorithm 1. Note that motifs have chronological order. That is, V_i appears earlier than V_j, if $i<j$. As Figure 2 shows, we obtained reconstructed series by stitching the motif sequence in order.

C. CONVOLUTION ON RECONSTRUCTED SERIES

Convolutional neural network model is firstly proposed by [34] and has gained noticeable performance in ImageNet classification task [23] in recent years. It is a specialized neural network that processes data with similar grid structure. However, recent research [12], [13] shows that CNN can also be comparative with RNN model while having faster calculation efficiency, which inspires us to select CNN model for sequence feature extraction.

By employing CNN module, a feature map is obtained by repeatedly applying the same function across sub-sequence

of the entire series. To be more specific, the feature maps are calculated by convolution operation of the input series with a linear filter, adding a bias term and then applying a non-linear function. If we denote the k-th feature map at a given layer as h^k, whose filters are determined by the weights W^k and bias b_k, then the feature map h^k is obtained as follows (for sigmoid non-linearity):

$$h_{ij}^k = sigmoid((W^k * x)_{ij} + b_k) \tag{1}$$

The Model is illustrated in Figure 1, where C_k represent k convolution operation and f_k represent k feature map. A convolutional layer is applied using k convolutional kernel with a sliding window of length w over input motif series to capture all local features. Since in this work we only consider one-dimensional price series, we adopt one-dimensional convolutional neural network. Before getting feature maps, we use sigmoid function as our non-linear activation function to improve the non-linear representation of the model:

$$\delta(\alpha) = \frac{1}{1 + e^{-\alpha}} \tag{2}$$

where α is the input signal. For the nonlinear transformed data, we use one more convolutional layer same as the previous one to extract more high-order feature form the k feature map obtained without pooling operation.

D. OPTIMIZATION AND PARAMETER SETTINGS

To train the above machine learning model, we need to introduce the objective function. Here we formalize the entire process of model and ultimate goal by minimizing the loss function as mean squared error:

$$L(y_i, f(X_{i:i+w}; W)) = \frac{1}{M}\sum_{i=1}^{M}(f(X_{i:i+w}; W) - y_i)^2, \tag{3}$$

where f represents the mapping or operation of the model, $f(X_{i:i+w}; W)$ is the predicted value and y_i is the true labeled value corresponds to sub-sequence series $X_{i:i+w}$ where W is the model parameter, M is the number of training samples.

At motif discovery and sequence reconstruction period, we set max vocabulary number $N = 5$, window size of time series $l = 30$, according to the results from recent research [25] that shows short period (a week or month) will get better performance on finance trending prediction task instead of long period. At the stage of feature extraction with CNN, two continuous 1D-convolution layers are used with the number of convolutional kernel $k = 64$ and each kernel size $s_k = 5$, strides $s_d = 1$. Besides a fully-connected layer with 128 output, the dropout layer is added to prevent overfitting and the drop rate is set to be 0.5. The final output is produced by a fully-connected layer with one output. Note that all these layers use sigmoid activation function. In the training period, we select stochastic optimization (*Adam*) as model's optimizer to minimize loss function. For each training phase, We update model parameters by batch iteration where batch size is 32 and maximum number of iterations is 500.

Algorithm 2 Motif-Based Financial Time Series Trend Forecasting

Input: Historical Time series $X_{i:i+n}$, N: max vocabulary number, w: sliding window size, y_i: label data of series X_{i+w+1}

Output: Probability P_{i+w+1} of ups and downs for X_{i+w+1}

1: $V = motif\ extraction(X_{i:w+1})$
2: $x_i = [V_1, V_2, ..., V_j], \quad j \leq N$
3: $loss \leftarrow min\ (f(x_i) - y_i)^2$
4: Return P_{i+w+1}

TABLE 2. The lengths of the data in each period of data analysis. Each time series is divide into fragments using sliding windows. Specifically, every fragment with length *l* is binarized into '10' sequence (corresponding to up and down, respectively) and taken as the input of the algorithm. In this paper, all of these fragments are splitted into training, validation and test.

Data	training	validation	test
SP	796	171	171
APPL	431	93	93
GE	653	140	140
GOOGL	152	33	33
BA	653	140	140
IBM	653	140	140
DIS	653	140	140
GT	560	121	120

IV. EXPERIMENTS

To explore whether it is possible to learn useful information from consecutive l days' price series before predicting the up and down of the $l + 1$-th day, comparing to the market price on the l-th day, we conduct extensive experiments on the financial datasets and compare the proposed method with several baseline methods.

A. EXPERIMENTAL DATA

Data description as Table 2 shows, we use daily financial time series from January 2 1962 to September 15 2017 obtained from the Yahoo Finance Website [1]. The datasets include the Standard & Poor's 500 stock ($S\&P\,500$) index and its individual stocks such as: Google Inc.(GOOGL), The Boeing Company(BA), International Business Machines Corporation(IBM), The Walt Disney Company(DIS), The Goodyear Tire & Rubber Company(GT), Apple Inc.(APPL), General Electric Company(GE). Each financial time series features a number of variables: *Open*, *High*, *Low*, *Close*, *Adjusted Close*, and *Volume*. In our experiments, we use the *Close* variable as our research object. Figure 3 shows the price variations in time for six examples. For each stock price series, the true label for each day's up and down is marked according to the following rule:

$$y_i = \begin{cases} 1 & X_i \leq X_{i+1} \\ 0 & X_i > X_{i+1} \end{cases} \tag{4}$$

[1] https://finance.yahoo.com/

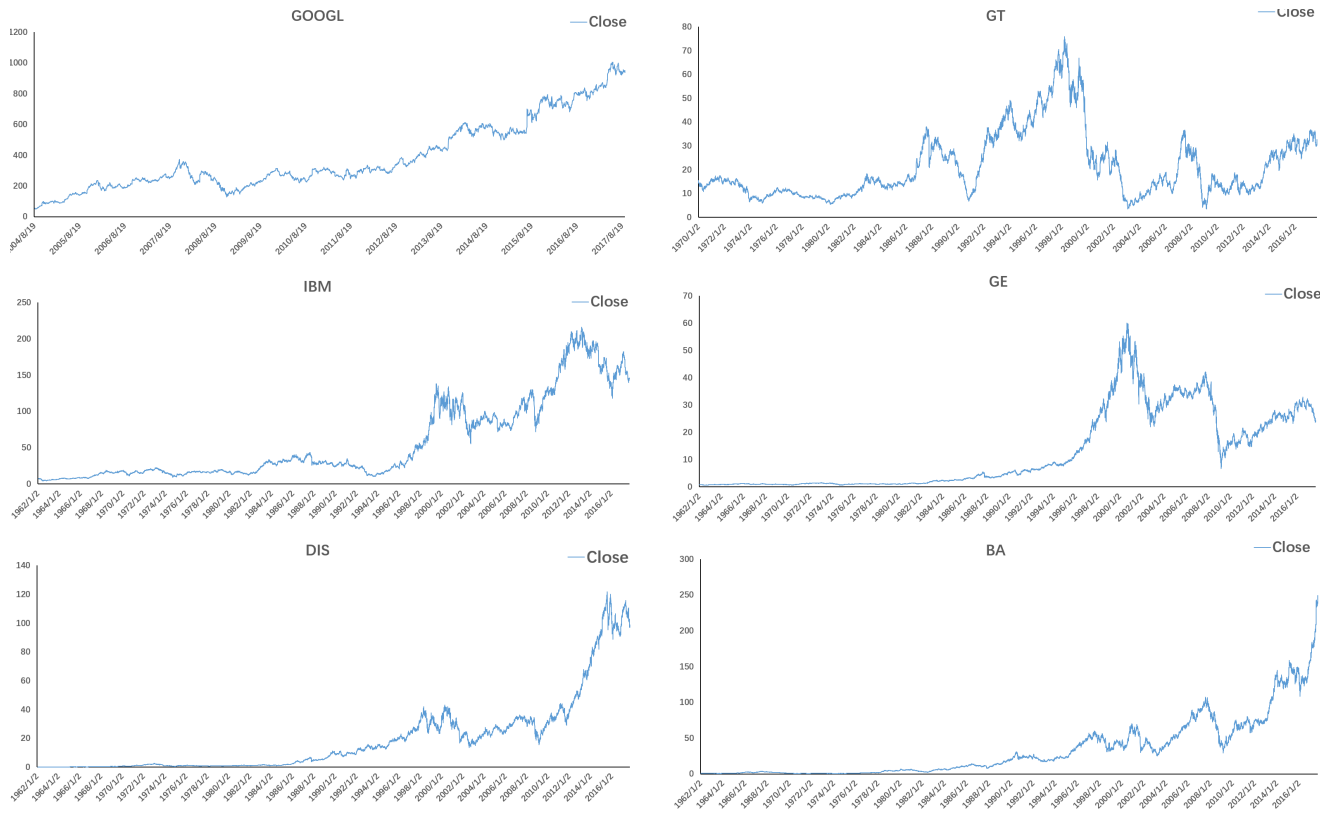

FIGURE 3. Example of six stock's close price.

TABLE 3. Confusion matrix of classification results.

True value	Forecast result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

where y_i denotes the up or down on i-th day, X_i is *Closed* value of each stock on i-th day and X_{i+1} is *Closed* value of each stock on i+1-th day.

B. EVALUATION

Generally, the prediction of the trend of stock price can be regarded as a two-category problem. Therefore, the prediction results fall into one of four cases based on the consistency of their real class labels and the predicted labels as true positive, false positive, true negative, or false negative. Let denote by TP, FP, TN, and FN the number of corresponding samples, respectively, the confusion matrix for the classification result has the following form:

We use the standard measure of Accuracy (Acc), $Recall$, $Precision$, F_1 score to evaluate the performance of individual stock prediction. These scores are calculated as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision}, \tag{8}$$

where TP is the number of samples in which the positive class sample is predicted to be a positive class, and TN is the number of samples in which the negative class sample is predicted to be a negative class, the denominator is the number of all test sample.

C. BASELINES

We compare our method to several baseline methods which are the state-of-the-art in different financial time series prediction tasks:

ARIMA [14]

: make a linear non-stationary sequence to become a smooth hybrid autoregressive moving average model(ARMA) by d-order differentiation.

Wavelet [21]

: combine the Stationary Wavelet Transform (SWT) with ARIMA method for time series forecasting.

HMM [18]: use Hidden Markov Model (HMM) to train on the past dataset of financial data and the trained HMM is used to search for the variable of interest behavioral data pattern from the past dataset.

LSTM [37]: apply a Long-Short term neural network to model S&P 500 volatility and predict stock price trend.

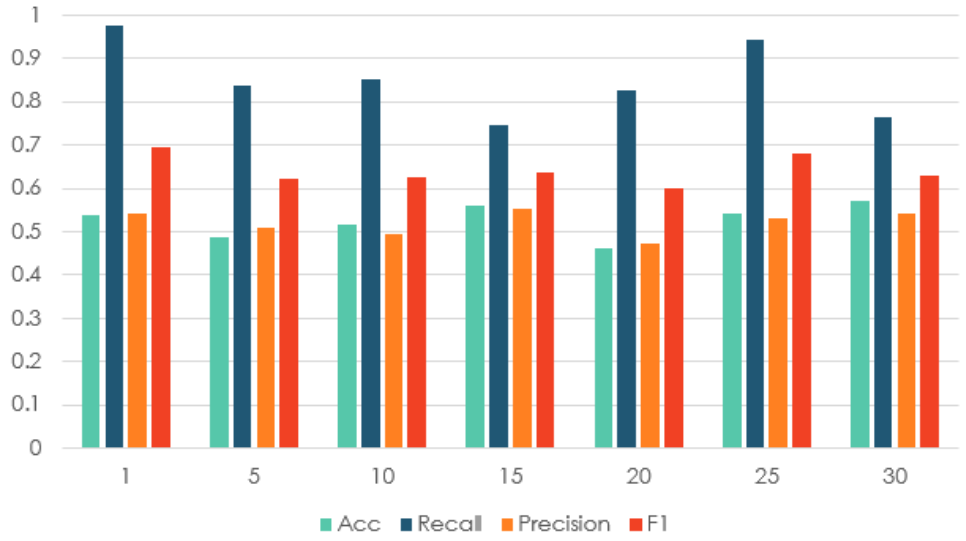

FIGURE 4. The number of various sliding windows size setting effect on S&P 500 data.

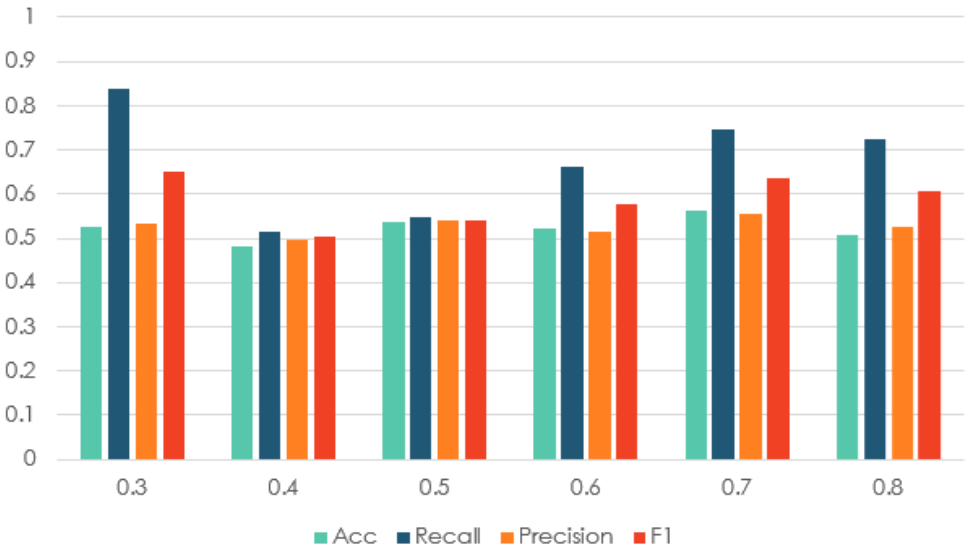

FIGURE 5. Different division ratio effect on S&P 500 in four evaluation criterion.

SFM [33]: a novel geometry-aware recurrent network to capture the multi-frequency trading patterns from past market data to make long and short term predictions over time.

D. RESULTS

By extracting trend features from the reconstructed sequence, we can see from Table 4 that our method improves the accuracy of the prediction results on almost all the datasets. From Table 4 and Table 5 we can see that LSTM is a competitive baseline considering its higher recall and precision scores on some datasets. However, LSTM is essentially a recurrent neural network model, which is time consuming than CNN in the training period due to the intrinsic vanishing gradient problem. It is also shown in both Table 4 and Table 5 that, traditional time series analysis methods and signal processing approaches (e.g., ARIMA and Wavelet) are inferior to machine learning based methods (e.g., HMM), which are $1\% - 3\%$ less in prediction accuracy and more less than deep learning model(LSTM, SFM). This reveals the limitations of traditional financial time series prediction methods on nonlinear times series modeling. On the whole, our model is superior to the baseline methods. In particular, with respect to accuracy and recall, it is higher than the traditional signal processing method with 6% to 7%, 6% higher than the machine learning methods and about 4% higher than the neural network models. Although f_1 scores do not achieve the best performance on all datasets, they are close to the best results.

We also visualize the prediction results in Figure 6. In Figure 6, the first row of each example represents the

TABLE 4. Comparison of our method to the baselines on the datasets using *accuracy*, *recall* scores.

Method	S&P 500		APPL		GE		GOOGL		BA		IBM		DIS		GT	
	acc	*recall*	acc	*recall*	acc	*recall*	acc	*recall*	acc	*recall*	acc	*recall*	acc	*recall*	acc	*recall*
ARIMA	50.88%	56.98%	46.24%	49.02%	53.90%	52.70%	48.47%	54.17%	50.35%	50.00%	46.81%	47.30%	51.77%	52.00%	54.54%	62.74%
Wavelet	48.77%	49.52%	51.83%	54.48%	51.74%	52.76%	51.80%	53.49%	51.74%	52.76%	51.18%	52.26%	51.43%	53.49%	49.67%	49.27%
HMM	51.53%	58.11%	50.02%	56.93%	50.47%	55.17%	49.96%	54.36%	50.32%	**55.36%**	49.46%	54.98%	50.49%	56.71%	50.05%	56.02%
LSTM	55.88%	71.91%	54.34%	**77.08%**	51.07%	**56.06%**	56.25%	**84.21%**	53.57%	39.42%	**62.85%**	63.34%	56.11%	27.53%	52.50%	39.21%
SFM	52.36%	51.85%	52.50%	52.18%	51.81%	50.32%	53.50%	54.66%	49.65%	48.92%	49.42%	48.50%	49.65%	48.92%	49.42%	48.50%
OUR	**56.14%**	**74.753%**	**58.06%**	60.47%	**56.43%**	41.38%	**63.63%**	71.45%	**54.29%**	54.78%	62.14%	**87.35%**	**58.57%**	**92.32%**	**56.20%**	**61.03%**

TABLE 5. Prediction on the datasets measured by *precision*, *f*1 scores.

Method	S&P 500		APPL		GE		GOOGL		BA		IBM		DIS		GT	
	precision	f_1	precision	f_1	precision	f_1	precision	f_1	precision	f_1	precision	f_1	precision	f_1	precision	f_1
ARIMA	51.04%	53.84%	51.02%	50.00%	56.52%	**54.54%**	47.70%	50.73%	52.85%	51.38%	49.29%	48.27%	54.92%	53.42%	47.06%	53.78%
Wavelet	52.41%	50.92%	53.55%	54.01%	54.04%	53.39%	49.46%	51.39%	**58.06%**	**55.28%**	53.09%	52.67%	55.67%	54.56%	**55.14%**	52.03%
HMM	54.56%	56.28%	52.75%	54.76%	53.04%	54.09%	51.73%	53.01%	52.84%	54.07%	50.79%	52.80%	53.12%	54.85%	53.23%	**54.59%**
LSTM	**56.14%**	63.05%	**54.41%**	**63.79%**	48.68%	52.11%	59.25%	69.56%	50.00%	44.08%	**68.80%**	58.60%	**63.33%**	38.38%	43.47%	41.23%
SFM	52.36%	51.85%	52.50%	52.18%	51.81%	50.32%	53.50%	54.66%	49.65%	48.92%	49.42%	48.50%	49.65%	48.92%	49.42%	48.50%
OUR	55.44%	**63.67%**	54.17%	57.14%	**74.05%**	52.87%	**67.87%**	**69.62%**	51.79%	52.86%	62.62%	**72.24%**	59.21%	**72.01%**	49.25%	54.02%

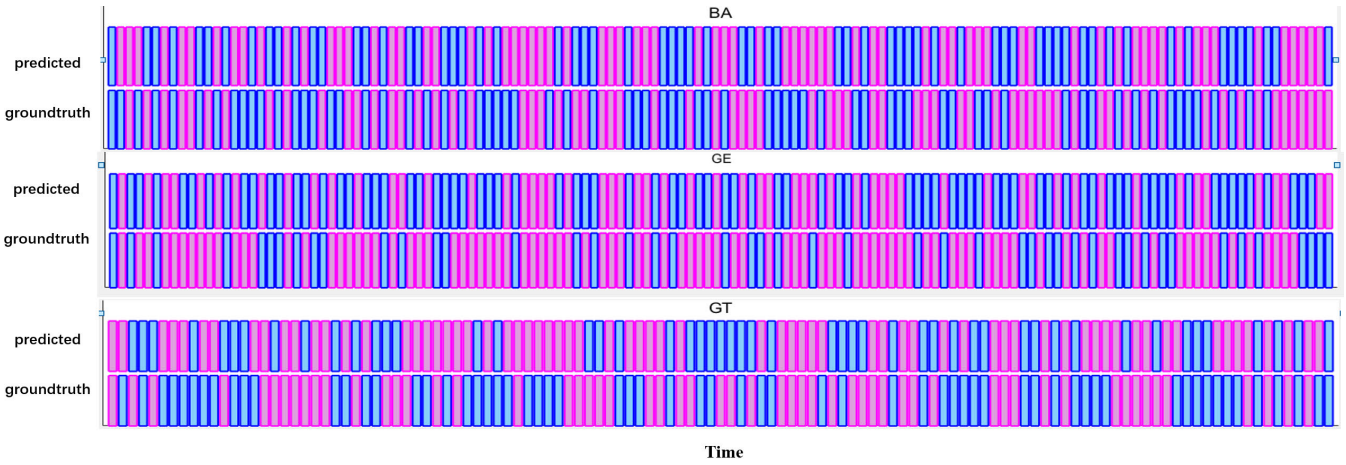

FIGURE 6. The prediction results by using the proposed method on BA, GT and GE datasets, respectively. Blue bars indicate downs and red bars represent ups.

predicted value, and the second row represents the true tag value, where red box indicates the down and blue box indicates the up. Each column in the instances corresponds to the actual value and the predicted value at a specific moment. The prediction is correct if two bars in the same column have the same color, otherwise the prediction is incorrect.

From the results we can conclude that the proposed method with sequence reconstruction by motifs has obvious advantages in the ups and downs trend prediction task compared to LSTM model and the existing frequency trading patterns modeling techniques for time series modeling, implying that the proposed model indeed extracts useful motifs from raw financial time series data, and the higher-order information obtained by the proposed CNN architecture facilitates the trend prediction. In fact, reconstructing time series using motifs can not only capture important features of the original data, but also filters out some low frequency interferential noises.

To explain the choice of model parameters in the experiment, we consider the impacts of the size of sliding windows w and sampling for training and test sets on S&P 500 trend prediction. As figure4 shows, the value of w influent model's performance heavily. When w equal to 1 or 25, the recall is very high while other metrics perform badly. It should be noted that in this task we concern both negative(downs) and positive(ups) sample because they both affect the effectiveness of the model and expect the model to have a good performance overall. Therefore, we select $w = 15$ as the value of sliding window size. Based on this, as figure 5(The x-axis represents the percentage of training sets and the rest are for validation and test. For example, 0.7 means the number of 70% data is used for training while the rest 15% is for validating and 15% is for testing) reflected that small training data will cause under-fitting because the pattern in the data are not effectively learned. At the same time, excessive training data with small quantity test data will effected by uneven

sample distribution. In this paper, we set 70% of each dataset are used for training and the rest of 30% for validating and testing. In this way, we treat each sequence segment and its label information as a sample pair and it's correspond data description shown in Table 2.

Besides, we further explore the reason of poor prediction performance on the BA dataset. We find that the time series are quite uneven, almost all prices vary towards up, while GOOGL and IBM have obvious fluctuation characteristics. The comparison of three experimental datasets shows that the proposed model succeeds in predicting frequently changed time series, while there still exists limitation for capturing anomalous patterns of time series. It should also be noted that considering the tradeoff between computational efficiency and validity of motif extraction, if sliding window w is too small, the motif extraction and sequence reconstruction will be greatly affected. However, if the window is too large, the calculation time of the motif extraction will be too long. So, in this paper, sliding window size w is set to be 15, window size of time series l is 30.

V. CONCLUSION

In this work, we introduce a novel technique for financial time series trend prediction by reconstructing time series via high-order structures, i.e., motifs. The underlying patterns in the reconstructed sequence are learned using convolutional neural networks, which provides useful information for ups and downs prediction. In comparison with the existing work that utilizes sequential model such as recurrent neural networks, our method is remarkably more efficient in terms of computational complexity. On the other hand, experimental results show the outperformance of our method on real financial time series datasets and validate the effectiveness on capturing trend information of stock shares. Our method sheds light on macroscopic pattern discovery in financial time series and provides a novel solution for price prediction.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive remarks and suggestions which helped us to improve the quality of the manuscript to a great extent.

REFERENCES

[1] E. Maiorino, F. M. Bianchi, L. Livi, A. Rizzi, and A. Sadeghian, "Data-driven detrending of nonstationary fractal time series with echo state networks," *Inf. Sci.*, vols. 382–383, pp. 359–373, Mar. 2017.

[2] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLoS ONE*, vol. 12, no. 7, 2017, Art. no. e0180944.

[3] F. Ye, Z. Liming, Z. Defu, F. Hamido, and G. Zhiguo, "A novel forecasting method based on multi-order fuzzy time series and technical analysis," *Inf. Sci.*, vols. 367–368, pp. 41–57, Nov. 2016.

[4] A. K. Rout, "Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 29, no. 4, pp. 536–552, 2017.

[5] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2018, pp. 95–104.

[6] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. J. Amaratunga, "Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting," *Appl. Soft Comput.*, vol. 54, pp. 246–255, May 2017.

[7] N. Nava, T. Di Matteo, and T. Aste, "Financial time series forecasting using empirical mode decomposition and support vector regression," *Risks*, vol. 6, no. 1, p. 7, 2018.

[8] L. B. Godfrey and M. S. Gashler, "Neural decomposition of time-series data for effective generalization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2973–2985, Jul. 2018.

[9] S. Jeon, B. Hong, and V. Chang, "Pattern graph tracking-based stock price prediction using big data," *Future Gener. Comput. Syst.*, vol. 80, pp. 171–187, Mar. 2018.

[10] X. Dai and M. Bikdash, "Trend analysis of fragmented time series for mHealth apps: Hypothesis testing based adaptive spline filtering method with importance weighting," *IEEE Access*, vol. 5, pp. 27767–27776, 2017.

[11] Z. Wang, F. Yang, D. W. C. Ho, and X. Liu, "Robust finite-horizon filtering for stochastic systems with missing measurements," *IEEE Signal Process. Lett.*, vol. 12, no. 6, pp. 437–440, Jun. 2005.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2014, pp. 1746–1751.

[13] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[15] F. E. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," *Omega*, vol. 29, no. 4, pp. 309–317, Aug. 2001.

[16] K.-J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, nos. 1–2, pp. 307–319, 2003.

[17] M. R. Hassan and B. Nath, "Stock market forecasting using hidden Markov model: A new approach," in *Proc. 5th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Sep. 2005, pp. 192–196.

[18] A. Gupta and B. Dhingra, "Stock market prediction using hidden Markov models," in *Proc. Students Conf. Eng. Syst. (SCES)*, Mar. 2012, pp. 1–4.

[19] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*. Berlin, Germany: Springer, 1981, pp. 366–381.

[20] C. Stolojescu, I. R. Cristina, S. Moga, P. Lenca, and A. Isar, "A wavelet based prediction method for time series," in *Proc. Int. Conf. Stochastic Modeling Techn. Data Anal.*, Chania, Greece, 2010.

[21] C. Stolojescu, A. Cusnir, S. Moga, and A. Isar, "Forecasting WiMAX BS traffic by statistical processing in the wavelet domain," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2009, pp. 1–4.

[22] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.

[23] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *Handbook Brain Theory Neural Netw.*, vol. 3361, no. 10, pp. 1–14, 1995.

[24] R. Xiong, E. P. Nichols, and Y. Shen. (2015). "Deep learning stock volatility with Google domestic trends." [Online]. Available: https://arxiv.org/abs/1512.04916

[25] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. IJCAI*, 2015.

[26] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using structured events to predict stock price movement: An empirical investigation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1415–1425.

[27] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Knowledge-driven event embedding for stock prediction," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 2133–2142.

[28] X. Xu, J. Zhang, and M. Small, "Superfamily phenomena and motifs of networks induced from time series," *Proc. Nat. Acad. Sci.*, vol. 105, no. 50, pp. 19601–19605, 2008.

[29] D. Yankov, E. Keogh, J. Medina, B. Chiu, and V. Zordan, "Detecting time series motifs under uniform scaling," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 844–853.

[30] A. Vahdatpour, N. Amini, and M. Sarrafzadeh, "Toward unsupervised activity discovery using multi dimensional motif detection in time series," in *Proc. IJCAI*, vol. 9, 2009.

[31] B. Hooi, S. Liu, A. Smailagic, and C. Faloutsos, "BeatLex: Summarizing and forecasting time series with patterns," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2017, pp. 3–19.
[32] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, "Exact discovery of time series motifs," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 473–484.
[33] L. Zhang, C. Aggarwal, and G.-J. Qi, "Stock price prediction via discovering multi-frequency trading patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 2141–2149.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[37] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 120, no. 2, pp. 654–669, 2017.
[38] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.
[39] Z.-K. Gao, S. Li, W.-D. Dang, Y.-X. Yang, Y. Do, and C. Grebogi, "Wavelet multiresolution complex network for analyzing multivariate nonlinear time series," *Int. J. Bifurcation Chaos*, vol. 27, no. 8, Jul. 2017, Art. no. 1750123.
[40] Z.-K. Gao, M. Small, and J. Kurths, "Complex network analysis of time series," *EPL (Europhys. Lett.)*, vol. 116, no. 5, 2017, Art. no. 50001.

MIN WEN is currently pursuing the degree in software engineering with Southwest Petroleum University, Chengdu, China. His research interests include time series analysis, machine learning, and data mining.

PING LI received the M.S. and Ph.D. degrees in computer science and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2004 and 2009, respectively. From 2009 to 2010, she was the Postdoctoral Researcher with the EIE Department, The Hong Kong Polytechnic University. From 2013 to 2014, she was with the Department of Computer Science, Tsinghua University. She is currently a tenured Professor with Southwest Petroleum University, Chengdu. Her current research interests include network science, machine learning, and their application in natural language processing.

LINGFEI ZHANG received the M.S. degree from the Department of Materials, University of Aalborg, Denmark. He is currently with the Corporate IT Department, Nomura Securities Corporation, Tokyo, Japan. His research interests include applied physics and financial time series.

YAN CHEN received the M.S. degree in computer science and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2006. She is currently an Associate Professor with Southwest Petroleum University, Chengdu. Her current research interests include machine learning, network science, and their application in intercross fields such as seismic data processing.