Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4024953, 12 pages
https://doi.org/10.1155/2022/4024953

Research Article

Stock Price Forecasting Based on Wavelet Filtering and Ensembled Machine Learning Model

Pengyue Wang,[1] Xuesheng Li,[1] Zhiliang Qin ,[1,2] Yuanyuan Qu,[1] and Zhongkai Zhang[1]

[1]*Weihai Beiyang Electrical Group Co, Ltd, Weihai, Shandong, China*
[2]*School of Mechanical, Electrical, and Information Engineering, Shandong University, Jinan, China*

Correspondence should be addressed to Zhiliang Qin; qinzhiliang@beiyang.com

Received 4 April 2022; Accepted 31 May 2022; Published 24 June 2022

Academic Editor: Muazzam Maqsood

Financial data are not only characterized by time-domain correlations but also heavily influenced by numerous market factors. In stock price analysis, the prediction of short-term movements is of much interest to investors and traders. In this paper, we consider forecasting price movements based on ensembled machine learning models, which is generally viewed as a challenging task due to noise components inherent in the data and uncertainties in various forms of financial information related to stock prices. To enhance the accuracy of trend predictions, we propose to use wavelet packet decomposition (WPD) and kernel-based smoothing techniques to remove high-frequency noise from the data, based on which we further perform feature engineering to obtain a comprehensive list of multidimensional technical features. Subsequently, we employ the light gradient boosting machine (lightGBM) algorithm to classify the change in the direction of the price trend that occurs in ten trading days. Numerical results on the Shanghai composite index show that the proposed approach has noticeable advantages over traditional statistical and machine learning methods when predicting near term price trends. Index terms—ensembled machine learning, feature correlation, financial data, LGBM, and wavelet denoising.

1. Introduction

Analyses on the stock market have received the attention of numerous traders and researchers. Specifically, the stock price forecast provides an important reference on setting a trading strategy or determining the appropriate timing for the transaction. Various theories and numerical techniques have been applied for decades to the stock market seeking to analyze the laws governing the price movements. Changes in the direction of price trends inevitably depend on a large number of factors, such as positive and negative news, company profiles, historical prices, and risk tolerances [1]. It is almost impossible to construct an all-compassing model incorporating the aforementioned factors. Moreover, the stock market itself includes transient and ubiquitous incidents involving individual companies and external incidents [2], such as diplomatic issues, and are impacted by random noises from market participants who hold different perspectives on economic outlooks. Needless to mention, personalized features, such as investors' sentiments, individual risk-bearing capacities, and even trading days or dates, also significantly affect the stock market, which makes the trend prediction a highly complex task [2].

Based on the generalized target, stock price prediction can be categorized into the tasks of classification and regression, respectively. The regression model obtains the estimated stock price directly, while the classification model produces the probability of the increment or the decrement over a certain time span. The resulting trading strategy is based on the rise or fall of the predicted price and hence provides traders with recommendations to buy or sell, respectively [3].

In the earlier stages of financial data analytics, researchers resorted to a number of statistical methods to materialize forecasting capacities [2], e.g., the support vector machine (SVM), (MLR), extra-trees algorithms (ET), autoregressive moving average (ARMA). In reference [1], Asghar et al. predicted closing prices on the Karachi Stock

Exchange (KSE)-100 index dataset based on a multilayered machine learning model. In reference [4], the authors used the ARMA model to make a forecast on the New York Stock Exchange (NYSE). In reference [5], the ARMA-GARCH algorithm was optimized, hoping to find a setting of near-optimal parameters to deliver the best returns for traders. However, due to the highly fluctuating and nonstationary nature of the stock market, statistics-based models are not effective in tackling the volatility and correlation structure in the data.

In recent years, the forecasting task significantly benefits from the development of machine learning approaches, which has led to notable advances in terms of numerical benchmarks. In reference [6], a support vector regression (SVR) algorithm is used to predict short-term returns. In reference [7], Nabipour et al. provide a comprehensive analysis and comparison of various models and evaluate their performance on the Tehran Stock Exchange (TSE). Experimental results show significant improvements on short-term return rates when machine learning models are used for binary classification tasks rather than performing regression on continuous data. In reference [8], a deep learning-based zero-inflated model is designed to conduct data analysis on financial data featured with irregularly spaced time. In reference [9], a fluctuation prediction model is presented to form trading strategies by processing a synthetic combination of online news, financial capacities, and social interest indicators.

While it is impossible to take into consideration all relevant multimodal data affecting the stock market, we believe that the impacts of these factors are manifested quantitatively as the numerical features of candlestick charts, which are also known as K-lines and typically used to represent both short-term and long-term fluctuations of stock prices. Hence, we construct a machine learning model by delving into empirical technical indicators and also deriving a set of signal characteristics as the model inputs. In this paper, we propose to use the state-of-the-art light gradient boosting machine (LGBM) model [10] to realize a binary classification task, i.e., to predict how the closing price in ten trading days changes over the corresponding value on the current date. The LGBM is an ensemble tree-based machine-learning framework and has critical superiorities over other models in that it takes advantage of sparsity characteristics of training data and is also viewed as an interpretable model [11, 12]. The novelties of the proposed approach in this paper are as follows.

(1) We use advanced signal processing methods, including wavelet filtering, to remove high-frequency noise components inherent in trading data and improve classification accuracies.

(2) We combine financial technical indicators with domain-specific signal characteristics to form a comprehensive list of features.

(3) Through numerical evaluations based on real-world market data, we demonstrate that the proposed approach achieves much higher accuracies over classical statistical models and conventional machine-learning models by resorting to effective data preprocessing and feature extraction techniques.

The rest of the paper is organized as follows: in Section 2, the proposed approach that incorporates wavelet filtering and feature engineering are presented. A brief description of the LGBM model is also included. In Section 2, we present numerical results on the Shanghai Stock Exchange (SSE). In Section 4, the conclusion is drawn.

2. Proposed Architecture

In this paper, we first present an introduction of the wavelet transform to preprocess stock data. Subsequently, we extract technical indicators to obtain representative information related to analyzing stock prices. Inspired by financial indicators originating from mechanical engineering, we obtain a set of features typically used for analyzing machine vibration signals. The extracted features are combined, normalized, and fed into a lightGBM model to make a binary classification on the rise or fall of the closing price over an interval of ten trading days.

2.1. Wavelet Denoising. Wavelet transform can be applied to remove high-frequency noise components from time sequences in signal processing. The set of wavelet functions [13] is derived from a wavelet function $h(t)$, which is further extended by $a = 2^m$, translated by $b = k * 2^m$, and normalized by [2]

$$h_{m,k}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) = \frac{1}{\sqrt{2^m}} h\left(2^{-m}t - k\right), \tag{1}$$

where m and k are defined by solving an expansion equation as indicated in [14, 15]. When the sequence $x(n)$ has $N = 2^s$ values, its expansion can be evaluated by [2]

$$x(n) = a_0 + \sum_{m=0}^{s-1} \sum_{k=0}^{2^{s-m-1}-1} a_{2^{s-m-1}+k} h\left(2^{-m}n - k\right). \tag{2}$$

In this paper, we choose the Haar wavelet as the wavelet function after extension numerical experiments, where the basis function $h_k(z)$ is defined as [2]

$$\begin{aligned} h_k(z) &= h_{pq}(z) \\ &= \frac{1}{\sqrt{N}} \begin{cases} 2^{p/2}, & \frac{(q-1)}{2^p} \le z < \frac{(q-0.5)}{2^p}, \\ 2^{p/2}, & \frac{(q-0.5)}{2^p} \le z < \frac{q}{2^p}, \\ 0, & \text{otherwise}, z \in [0,1], \end{cases} \end{aligned} \tag{3}$$

where $h_0(z) = h_{00}(z) = 1/\sqrt{n}, Z \in [0,1]$ and k is obtained by $k = 2p + q - 1$.

The original signal can be decomposed by the wavelet transform to obtain the approximate components and

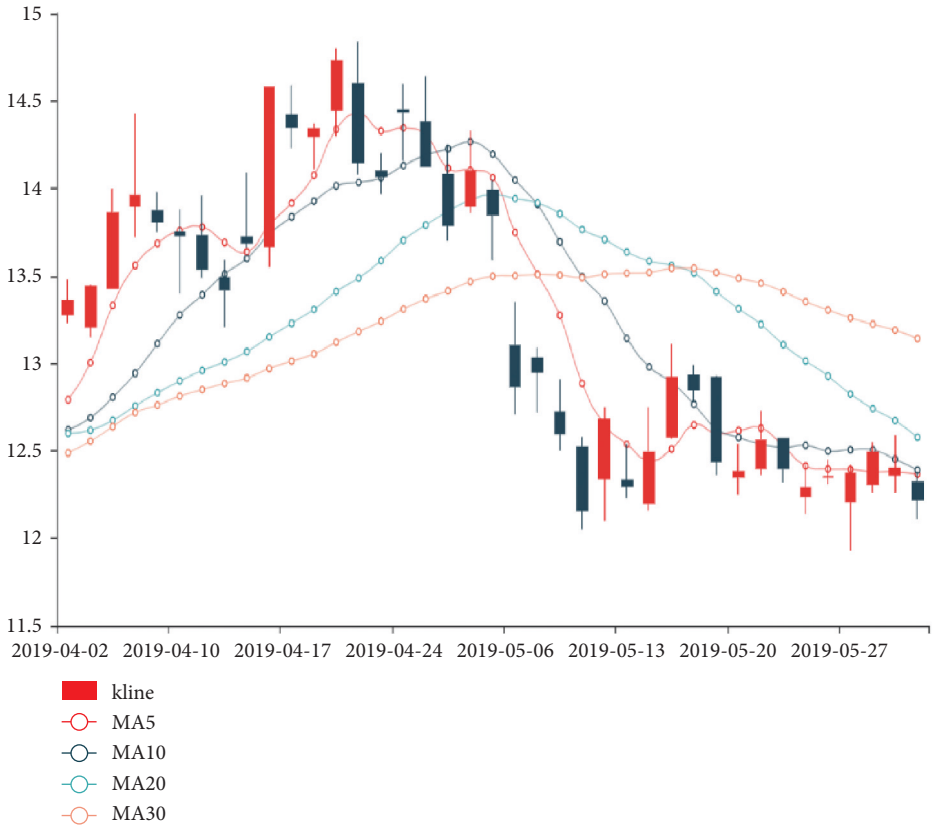


Figure 1: 000001.SZ candlestick graph.

detailed components, which are subsequently thresholded for the purpose of reconstructing a denoised version of the original signal. The decomposition procedure may be viewed as a multiresolution analysis [16, 17], and involves the following steps. First, we construct the transformation basis by using a scaling function and a wavelet function, which are defined by [18].

$$\begin{cases} \phi_{j,k}(t) = 2^{j/2}\phi\left(2^{j}t - k\right), \\ \psi_{j,k}(t) = 2^{j/2}\psi\left(2^{j}t - k\right), \end{cases} \tag{4}$$

where j denotes the dilation or the visibility in frequency and k specifies the position.

To obtain the wavelet decomposition of a signal, we need to ensure that the scaling function of the signal is orthogonal to its translated variant. Moreover, the subspaces that are obtained based on spanning the scaling function at low scales are required to be nested within those obtained at higher scales.

2.2. Financial Technical Indicators. Figure 1 shows the daily candlestick chart of the stock with the code 000001.SZ ranging from April to May 2019. The chart is a type of financial representation that shows the price action for an investment market. It consists of specific candlesticks that denote the opening, closing, and high and low prices each day over a given time interval, which makes it more useful than traditional lines that simply connect the dots of closing prices.

Moreover, the chart can be used for identifying trading patterns that help technical analysts to establish trading modes. From a pragmatic perspective, candlestick patterns can be formed by grouping two or more candlesticks in a certain fashion. The pattern trend provides an intuition to predict the direction of the price movements. In Figure 1, the rectangle part of the daily candlestick is called the real body showing the link between opening and closing price and represents the price gain or loss for the specified period. The thin lines above and below the real bodies are called shadows and also referred to wicks, which show the highest and lowest prices in the trading session. The filled red color of the real body means that the price is closed higher than its opening price; while the green color indicates that a session is closed lower. Although the candlestick can be interpreted using a variety of methods, the relationship between the opening and closing price is considered the most vital information on price movements. In particular, the close price is generally considered as the most important indicator to assist a trader in forming his short-term trading strategy.

Over the past years, researchers developed a large number of technical indicators based on the statistics of the candlestick chart to analyze the price fluctuation. The moving average convergence and divergence (MACD) indicator is deemed a most well-known trend-following momentum oscillator to represent quantitatively the relationship between the moving averages (MA) of the closing price. Mathematically, the standard MACD is calculated

based on the difference (DIF) between fast (typically 12-day) and slow (26-day) exponential moving average (EMA). Changes in the time periods used for the calculation can be made to accommodate a trader's specific targets or a particular type of trading. The EMA [19] is a type of moving average (MA) that places a larger weight on the most recent data points and reflects sensitively the near term price changes [20, 21].

The MACD histogram is a mathematical tool to evaluate the signed distance between the MACD and its signal line based on the 9-day EMA, which is also known as the divergence exponential average (DEA). The calculations of the MACD are given as follows [2]:

$$\begin{aligned} \mathrm{EMA}_n(x) &= \mathrm{EMA}_{n-1}(x) \times \frac{x-1}{x+1} + C_n \times \frac{2}{x+1}, \\ \mathrm{DIF}_n &= \mathrm{EMA}_n(x_{\mathrm{fast}}) - \mathrm{EMA}_n(x_{\mathrm{slow}}), \\ \mathrm{DEA}_n &= \mathrm{DEA}_{n-1} \times 0.8 + \mathrm{DIF}_n \times 0.2, \end{aligned} \tag{5}$$

where x is the number of days and C_n is the closing price on the nth day. Typically, traders use the MACD histogram to anticipate changes in the market momentum. For instance, for the positive values of DIF and DEA, the MACD line crosses the signal line to produce an uptrend divergence and output a buy suggestion. For negative DIF and DEA values, the signal line traverses the MACD line, which advises a sell recommendation based on the negative divergence behavior.

As the MACD alone may generate false predictions, experienced traders rely on complementary trend measurement indicators. A commonly used indicator is known as the KDJ index [21], which is otherwise known as the random index. It is a practical technical indicator that is commonly used in short-term trend analysis. It derived from the stochastic oscillator, which, however, differs from the latter by including an extra J line. Values of K and D lines show whether a stock is overbought or oversold; while the J line represents the divergence of the D line from the K line [22]. The indicator incorporates price levels accounting for the amplitude of fluctuations in the prices. The fastest, slowest, and medium indices K, D, and J are calculated as follows [21, 23]:

$$\begin{aligned} \mathrm{RSV}_n &= \frac{(C_n - L_n)}{(H_n - L_n)} \times 100, \\ K_n &= \frac{2}{3} \times K_{n-1} + \frac{1}{3} \times \mathrm{RSV}_n, \\ D_n &= \frac{2}{3} \times D_{n-1} + \frac{1}{3} \times K_n, \\ J &= 3D - 2K, \end{aligned} \tag{6}$$

where n denotes the nth trading day and C_n denotes the closing price on the nth day. Note that H_n and L_n denote the highest and lowest price within n days, respectively.

Another popular technical indicator is the relative strength index (RSI), which is typically employed in technical analysis to evaluate the magnitude of price changes.

Table 1: Mechanical signal indicators.

No.	Equation
1	$1/L \sum_{l=1}^{L} s(l)$
2	$\sqrt{\sum_{l=1}^{L} (s(l) - c_1)^2/L - 1}$
3	$\sqrt{\sum_{l=1}^{L} (s(l))^2/L}$
4	$1/L \sum_{l=1}^{L} s(l)$
5	$\left(\sqrt{\sum_{l=1}^{L} \sqrt{s(l)}/L}\right)^2$
6	$\sum_{l=1}^{L} (s(l) - c_1)^3/(L-1)c_2^3$
7	$\sum_{l=1}^{L} (s(l) - c_1)^4/(L-1)c_2^4$
8	$\max(\lvert s(l)\rvert)/c_4$
9	$\max(\lvert s(l)\rvert)/c_5$
10	c_3/c_4
11	$\sum_{k=1}^{K} f(k)/k$
12	$\sqrt{\sum_{k=1}^{K} (f(k) - c_{11})^2/K - 1}$
13	$\sum_{k=1}^{K} (f(k) - c_{11})^3/K(c_{12})^3$
14	$\sum_{k=1}^{K} (f(k) - c_{11})^4/K(c_{12})^4$
15	$\sum_{k=1}^{K} v_k f(k)/\sum_{k=1}^{K} f(k)$
16	$\sqrt{\sum_{k=1}^{K} (v_k - c_{15})^2 f(k)/K}$
17	$\sqrt{\sum_{k=1}^{K} v_k^4 f(k)/\sum_{k=1}^{K} v_k^2 f(k)}$
18	$\sum_{k=1}^{K} v_k^2 f(k)/\sqrt{\sum_{k=1}^{K} f(k) \sum_{k=1}^{K} v_k^4 f(k)}$
19	$\sum_{k=1}^{K} (v_k - c_{15})^3 f(k)/Kc_{16}^3$
20	$\sum_{k=1}^{K} (v_k - c_{15})^4 f(k)/Kc_{16}^4$

Hence, it is feasible to use this indicator to estimate if the trading condition is overbought or oversold. The RSI measures both the speed and the change rate in price movements. RSI values are typically estimated over a 14-day period and fluctuate between zero and 100 and can be obtained as [24]

$$\begin{aligned} RS &= \frac{1}{N}\left(\sum_{i=2}^{N}(Up_{t-i})\right) \div \frac{1}{N}\left(\sum_{i-2}^{N}(Dw_{t-i})\right), \\ \mathrm{RSI} &= 100 - \frac{100}{1 + RS}, \end{aligned} \tag{7}$$

where Up and D_w represent the upward and downward movements in terms of the closing price, respectively. Other indicators include the on-balanced volume (OBV) to describe changes in volume, the William % R to show the current closing price related to the high and low price of the past time period, and the price channels to identify an upward thrust to signal the start of an uptrend.

The calculation of the abovementioned five most popular financial indicators provides us with numerical metrics to quantitatively characterize the moving trend of stock prices.

Furthermore, inspired by reference [24] that effectively extended the mechanical engineering concept RSI to the field of financial data analysis, we perform extensive numerical evaluations of the features that are specifically used in the empirical modeling of mechanical vibration signals [25, 26] and select the following twenty indicators based on the criterion of performance optimization, as shown in Table 1. Each feature reflects an attribute of a time sequence since the vibration signal and the stock price can be both viewed as time sequences and bear much resemblance with each other. Certain mechanical features place more significance on time-domain properties such as magnitudes and energy differences; while the others indirectly represent frequency-domain properties in terms of the zero crossing rate and the position change of frequency bands.

Combining financial indicators with mechanical signal characteristics, we form a comprehensive list of extracted features as the inputs to the proposed model. As various features differ noticeably in magnitudes, we apply the min-max scaling to perform the data standardization for each feature.

$$y_j = \frac{x_i - \min_{1<j<n}\{x_j\}}{\max_{1<j<n}\{x_j\} - \min_{1<j<n}\{x_j\}}. \tag{8}$$

In the numerical experiments, a 30-day sliding window interval is used to form the model inputs and the corresponding binary targets. Considering that we have a two-dimensional array of extracted features for each target, we further extract four basic statistics, i.e., maximum, minimum, mean, and standard deviations of the input features as shown in (9) and (10) to reduce the model input to a one-dimension (1-D) vector for each target [27]

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{9}$$

$$D(x) = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - E(x)\right)^2. \tag{10}$$

In financial analysis, it is worth mentioning that the date is deemed an important feature. For instance, the date of the mutual fund redemption tends to result in market volatilities. To fully exploit this feature, we convert trading dates into monthly, weekly, and daily variables, respectively, by resorting to the one-hot encoding technique, thus turning these categorical variables into numerical vectors.

Finally, we have a tabular form of processed data, where each row corresponds to a number of derived features to the model, which are composed of technical indicators, mechanical characteristics, and date variables, as well as a binary target, to be predicted to signify the rise or fall of the closing price in ten trading days as compared with the current date.

2.3. LGBM Model. The decision tree model [28] seeks to maximize the differences between class probabilistic distributions. By building a tree structure that satisfies division conditions, samples are classified and predicted based on the optimized model. The gradient boosting decision tree (GBDT) algorithm is a family of lifting tree models seeking to improve the performance of the decision tree model by using the technique of classification and regression trees (CART). By initializing a weak classifier $f_0(x)$, the negative gradient of each sample can be obtained by [29]

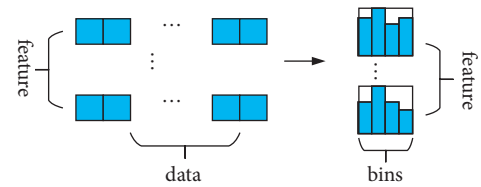


Figure 2: LGBM histogram-based discrete process.

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c). \tag{11}$$

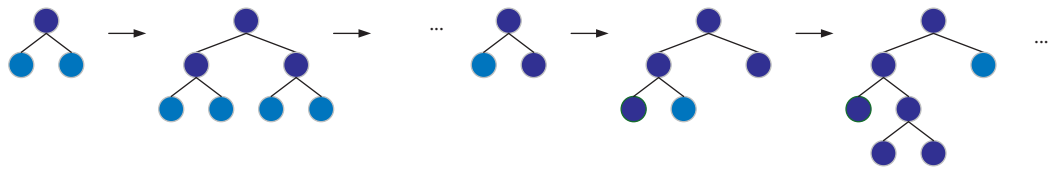
Hence, the obtained residual is used as the updated ground-truth value of the sample, and the training data can be updated for the next decision tree. Following this procedure, the final learner is obtained by calculating the best fitting value for the leaf area and repeatedly updating the strong learner as follows:

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M}\sum_{j=1}^{J} \gamma_{jm} I\left(x \in R_{jm}\right). \tag{12}$$

In order to find a suitable split point, the GBDT algorithm needs to scan all data subsets. Hence, it has an excessively slow computing speed and incurs a large amount of memory. To overcome these limitations, an improved memory-efficient version, i.e., the gradient boosting machine (GBM) algorithm is proposed in [10].

Despite the popularity of deep learning in recent years, the GBM algorithm generally performs better in the tasks of analyzing tabular data. In many scenarios, this algorithm is preferred in practical implementations due to its interpretability, fast convergence, and possibility to incorporate modularly domain-specific prior knowledge. It uses the additive models of weak learners to optimize a specific loss function and fine-tune hyper-parameters based on the gradient descent algorithm. Specifically, there are two categories of powerful GBM algorithms, i.e., the extreme GBM (XGB) and the LGBM models. Both algorithms obtain performance comparable to deep-learning-based convolutional neural networks (CNN) in data analytic tasks.

In the training process, the XGB algorithm traverses the dataset multiple and generally displays a much slower convergence behavior [30]. On the contrary, the LGBM model distributes computations across multiple nodes and employs a parallelized hierarchical learning approach to derive inherent patterns from large-scale data. For self-containing purposes, we include the abovementioned process in Figure 2. Specifically, the lightGBM algorithm performs the evaluation on a subset of training data to obtain an



(a) (b)

Figure 3: Tree growth: (a) level-wise growth and (b) leaf-wise growth.

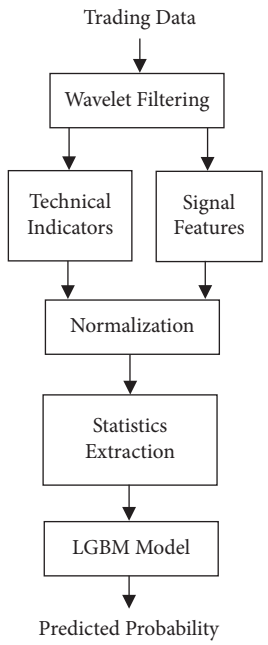


Figure 4: Flowchart of the proposed model for stock price forecasting.

entropy metric based on which a nearly optimal segmentation can be made. Hence, it effectively achieves a reduction in terms of memory usage, communication costs, and the computational resources needed to obtain gains for tree-splits. The theory shows that this method does not suffer from the loss of the accuracy and is capable of achieving good performance with large datasets while with a significant reduction in training time as compared with the XGB algorithm.

Figure 3 shows that the construction of the LGBM follows a leaf-wise approach, reducing more training losses than the conventional level-wise algorithms [30]. When growing on an equivalent leaf, the leaf-wise algorithm optimizes the target function more efficiently than the level-wise algorithm and leads to better classification accuracies, which can rarely be achieved by other boosting algorithms. In this paper, we use the LGBM model incorporating a depth-limited leaf growth strategy in numerical experiments. A comparison with the XGB algorithm is also made by evaluating their accuracies on typical stocks across four industry sectors.

Figure 4 shows the schematic flowchart of the proposed approach. The algorithm consists of the wavelet filtering module to denoise the raw data.

Table 2: Calculation of technical indicators.

	Original	Mean30	Max30	Min30	Std30
Open	2.94	−3.62	2.94	−1.73	0.99
High	4.43	4.22	4.46	4.07	0.08
Low	4.33	4.13	4.33	3.99	0.07
Close	4.36	4.175	4.4	4.02	0.09
Volume	124403	119592	321326	60110	50861
MACD	0.015	0.097	0.015	0.199	0.068
K	79.31	48.25	81.94	13.65	23.66
D	78.45	46.65	78.45	14.89	21.99
J	81.03	51.44	109.27	-5.53	33.57
OBV	1539942	1086483	1664345	727524	218032
RSI	67.66	50.08	77.74	23.26	14.92
MA5	4.314	4.162	4.314	4.054	0.061
MA10	4.256	4.163	4.256	4.094	0.05
MA20	4.215	4.185	4.358	4.123	0.061
MA30	4.175	4.286	4.702	4.152	0.167

Table 3: Stock forecasting for coal industry.

Codes	Industry	Accuracy (%)
000552.SZ	Coal	70.8
000933.SZ	Coal	66.6
000937.SZ	Coal	75.5
002128.SZ	Coal	68.7
600381.SH	Coal	72.9
600714.SH	Coal	74.1

Based on the filtered sequence, we proceed to derive empirically validated financial technical indicators including MACD, KDJ, RSI, and OBV. Furthermore, we extract a number of mechanical features typically used in the analysis of machine vibration signals to reflect both time-domain and frequency-domain properties. To achieve numerical stabilities in experiments, we perform the standardization of the derived features over a sliding window of 30 trading days and normalize all values to the range of 0 to 1.0 based on the min-max scaling. The target of the prediction is formed by calculating the difference between the current close price and the closing price obtained over the period of ten trading days, and thus generating a positive or negative binary label depending on the rise or fall of the trend. Finally, we apply the random-search technique to optimize the LGBM parameters and use the optimized model to predict the trend of the closing price. After an extensive search for optimized parameters, we set the initial learning rate to 0.05, the number of leaves to 120, and the maximum depth of the LGBM tree to 6.

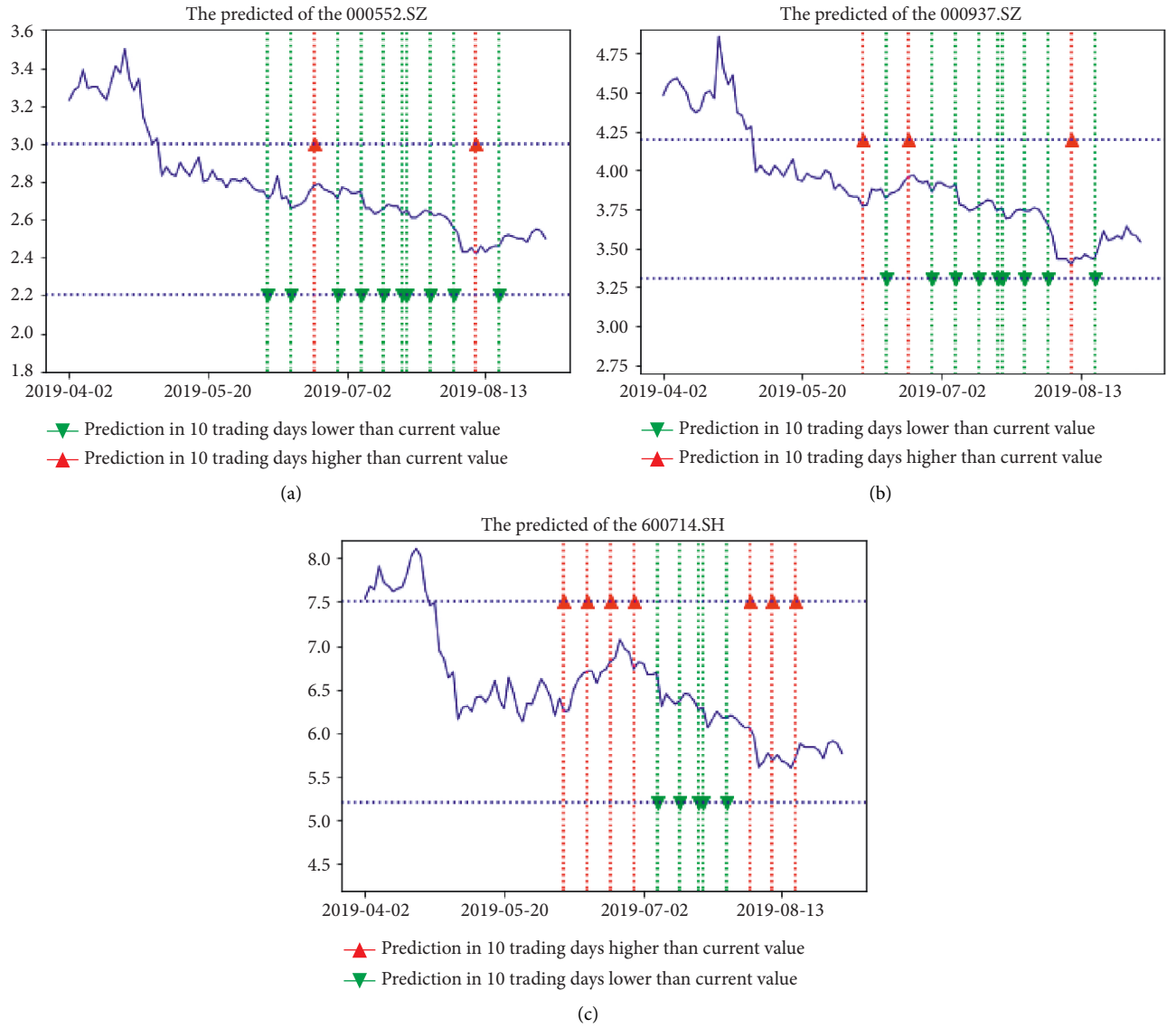


Figure 5: Predictions of stocks in the coal industry: (a) 000552.SZ; (b) 000937.SZ; (c) 600714.SH.

3. Numerical Experiments

We use Tushare [31] to retrieve stock price datasets across the real estate, coal, electric power, and cement segments in the Shanghai Stock Market. Tushare is a convenient tool to perform data retrieval, cleansing, and storage of financial data due to its simple application interface (API) and short response time. Moreover, it is equipped with a set of readily used visualization functions to check the stock price data, which are often susceptible to the existence of data errors or outliers.

A typical metric used to perform the evaluation on classification models is the confusion matrix, which shows the errors (i.e., confusions) among different classes. The results of correct classifications are displayed on the diagonal of the matrix, while incorrect results are expressed as off-diagonal entries. Based on the confusion matrix, we further calculate a simple metric, i.e., accuracy, as the percentage of correct predictions out of the total number of samples. Accuracy is recognized as the most widely used empirical metric in the literature. Hence, we use it in this paper to benchmark the performance of various algorithms.

Table 4: Stock forecasting for real-estate industry.

Securities code	Industry	Accuracy (%)
600383.SH	Real estate	91.7
000014.SZ	Real estate	83.3
600095.SH	Real estate	74.7
600159.SH	Real estate	72.9
600322.SH	Real estate	75.2
000011.SZ	Real estate	68.7

Of utmost importance in the task of trend predictions is eliminating any possibility of the data leakage, i.e., using validation data in the training procedure and thus resulting in an unreasonable though attractive high accuracy. To

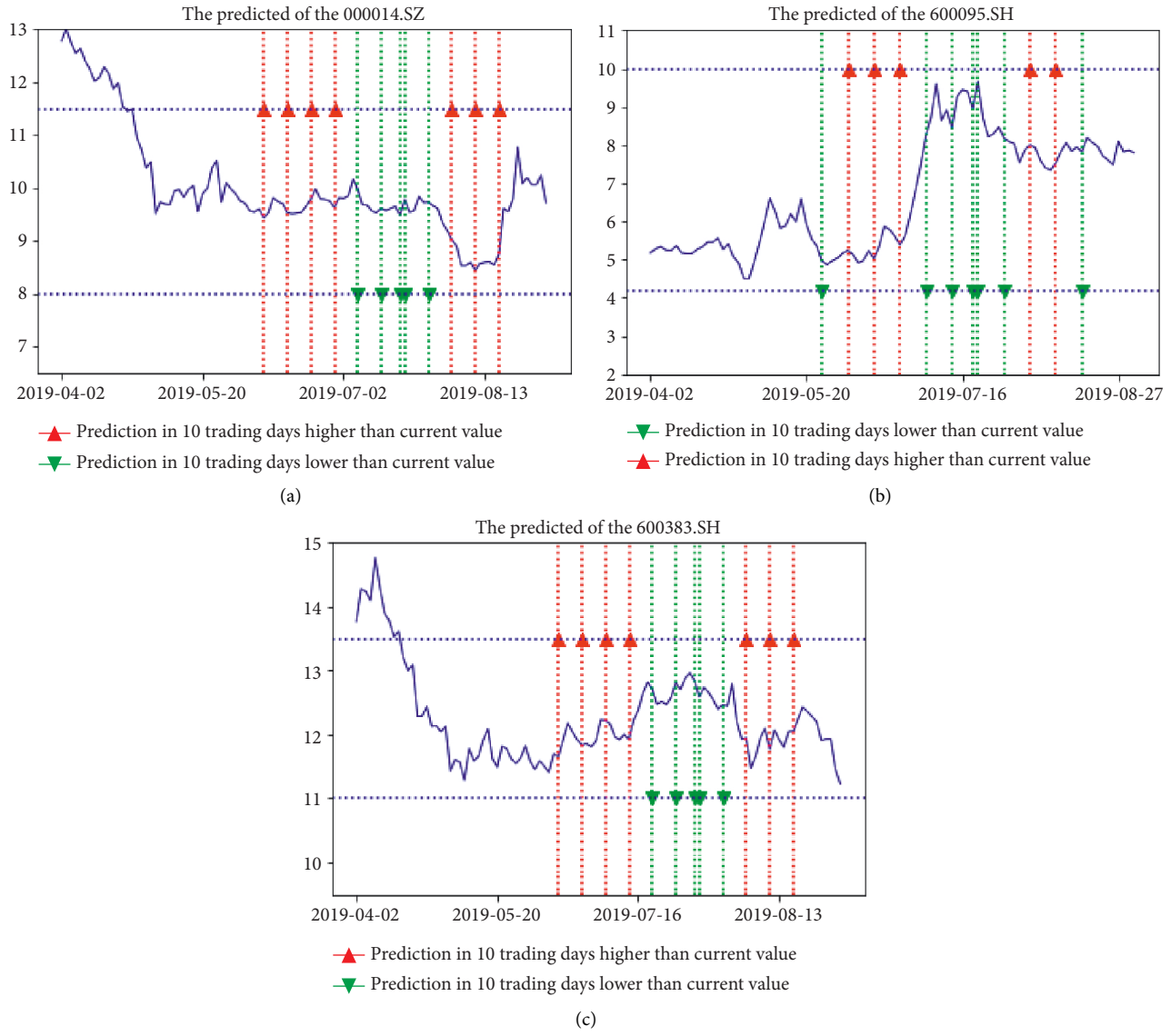


Figure 6: Predictions of stocks in the real estate industry: (a) 000014.SZ; (b) 600095.SH; (c) 600038.SH.

Table 5: Stock forecasting for the electric-power industry.

Codes	Industry	Accuracy (%)
000543.SZ	Electric power	66.6
000767.SZ	Electric power	81.2
002479.SZ	Electric power	83.7
600098.SH	Electric power	75.0
600131.SH	Electric power	77.1
600509.SH	Electric power	72.9

Table 6: Stock forecasting for the cement industry.

Codes	Industry	Accuracy (%)
000546.SZ	Cement	74.3
000401.SZ	Cement	71.4
002233.SZ	Cement	83.3
600449.SH	Cement	58.3
600881.SH	Cement	81.2
000877.SZ	Cement	75.3

ensure a fair comparison with other algorithms, we form a strictly non-overlapping subset of train data and validation data. That is, the data ranging from January 2015 to January 2019, is used to train the model; while validation is conducted on the financial data over the interval between February 2019 to September 2019 [23].

For illustration purposes, Table 2 shows the calculation of certain financial indicators by taking a sliding window over 30 trading days. It is noted that various indicators have a vastly different and dynamic range, which implies that normalization is a necessary operation to ensure the stability of the model in the training process and also enable the model to learn discriminant features across various domains.

To show quantitative results, Table 3 shows the accuracies of the predicted trends by the proposed method,

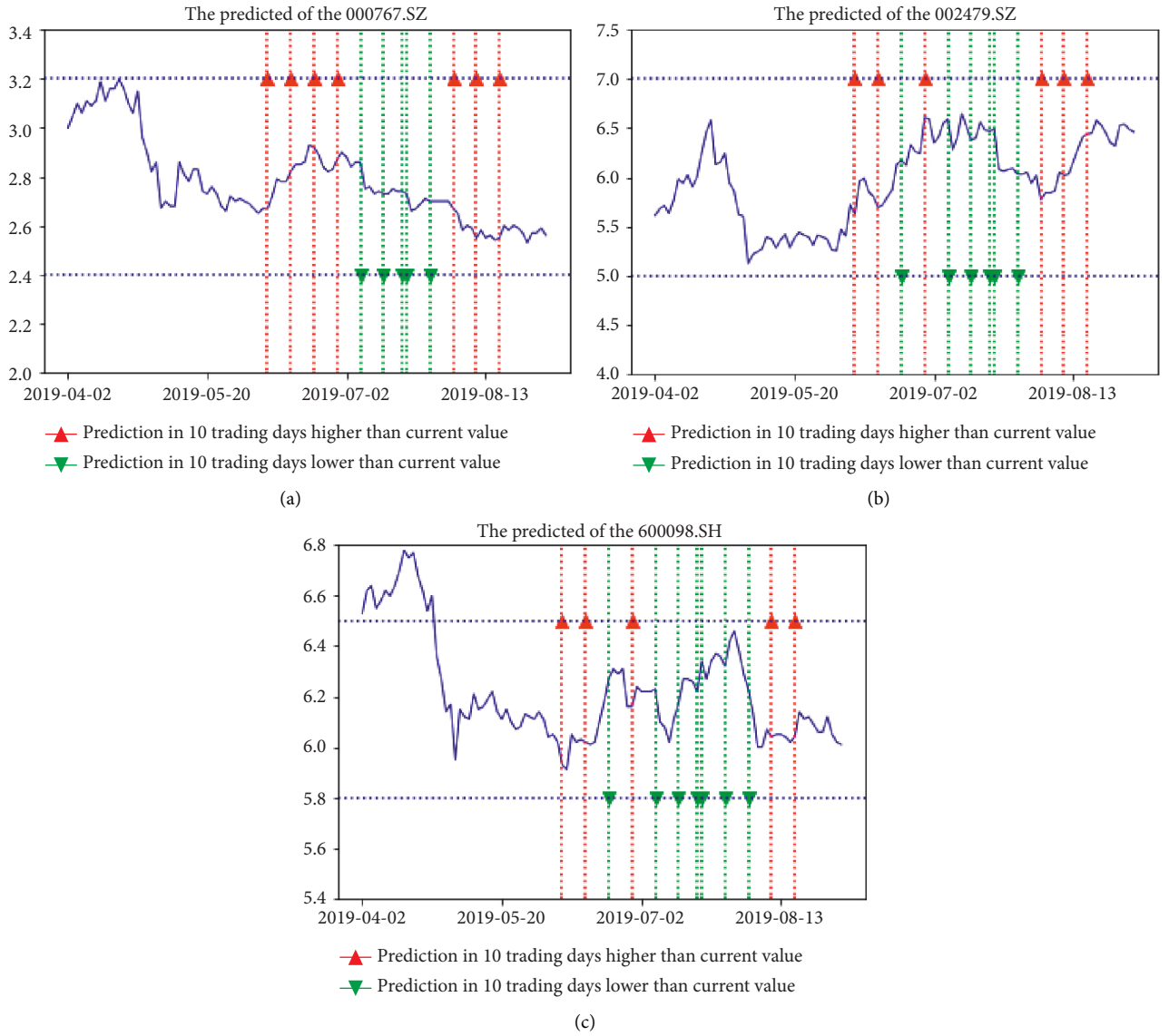


Figure 7: Predictions of stocks in the electric power industry: (a) 000767.SZ; (b) 002479.SZ; (c) 600098.SH.

which are numerically evaluated on a number of stocks belonging to the coal industry.

Table 3 shows that the proposed approach obtains an accuracy of nearly 70% for most companies in the coal industry by employing financial technical indicators and signal-domain features. The accuracy is considered to be impressive for a short-term prediction task. In Figure 5, we graphically present the prediction results of several stocks, e.g., 000552.SZ, 000937.SZ, and 600714.SH for illustration purposes, where colored triangles denote the trend of the closing price. The date on which triangles are drawn represent the current trading date. Specifically, red triangles denote an increment that tends to occur over the specified interval; while green triangles denote a predicted decline. In Table 4, we present the predictions over the real estate industry. It is of interest to observe that the proposed approach obtains better results when compared with the coal industry. Figure 6 shows a visual presentation of several stocks. It is shown that the trends are predicted accurately on most trading dates prior to abrupt changes in the curve.

Similarly, Tables 5 and 6 show the prediction accuracies over the electric power and the cement industry, while the graphs of individual stocks are presented in Figures 7 and 8, respectively. By presenting the numerical results of the proposed model across four industries, we have established a benchmark to compare the performance with other models including the conventional SVM and RF models as well as the powerful XGB model. A comparison with deep-learning CNN and attention-based transformer models would be considered as future work.

On the same test set, we perform classification based on the SVM, random forest (RF) [32], ARMA, and autoregressive integrated moving average model (ARIMA). For a fair comparison, we have generated the same set of features when compared with the conventional machine-learning models such as SVM and RF. For statistical models, note that

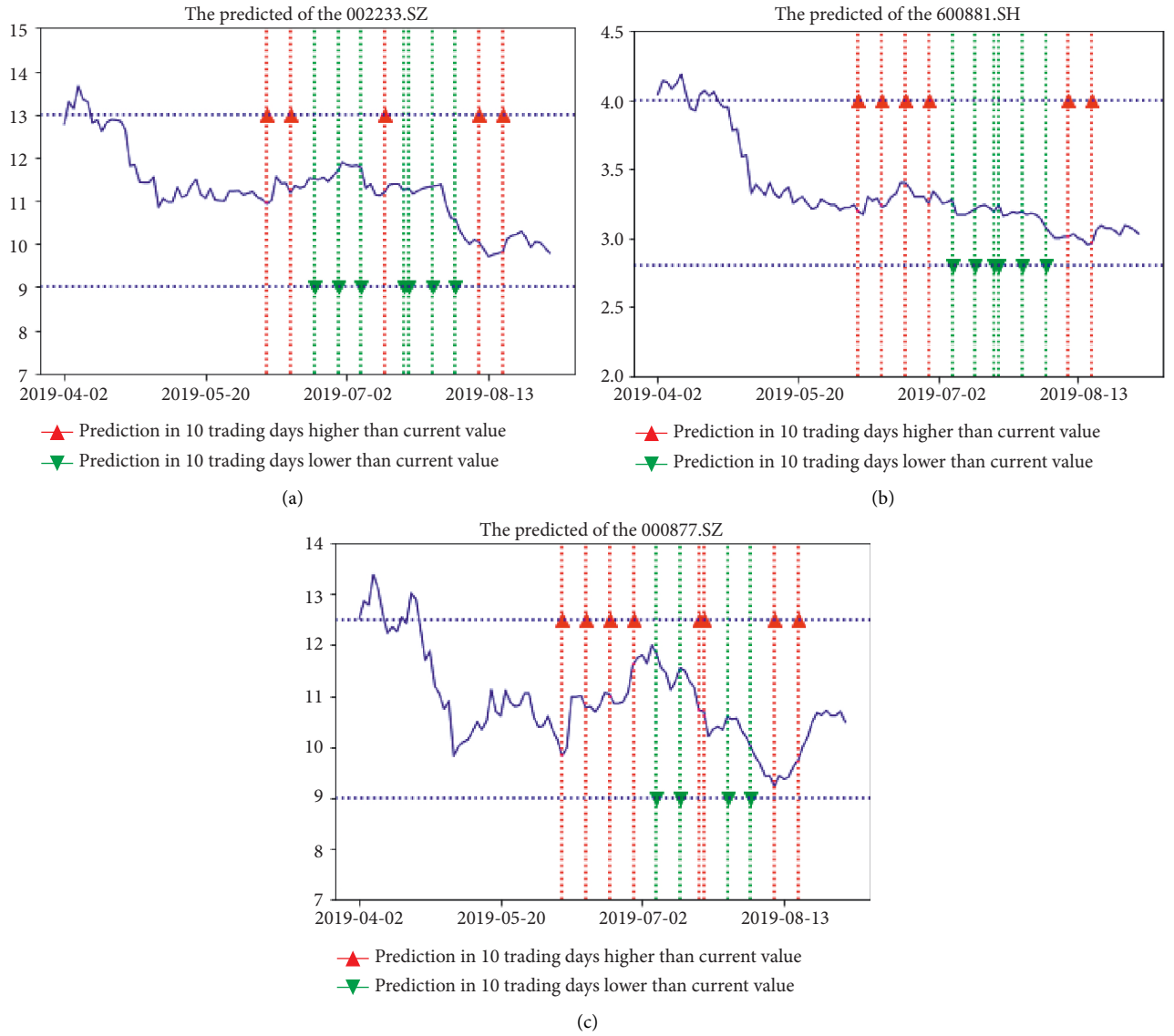


Figure 8: Predictions of stocks in the cement industry: (a) 002233.SZ; (b) 600881.SH; (c) 000877.SZ.

we have to access the test set so as to predict the trend of closing prices due to the requirements of the models. Table 7 shows the predicted accuracies on the test set across four industries. We have included various ensembled models in the comparisons, i.e., the RF model that is composed of random trees and an ensembling approach to average the prediction probabilities of the RF and the SVM models. The XGB model [33], which is generally viewed as a powerful ensembled learning algorithm, is also evaluated on the validation set. It is shown that the proposed LGBM model based on a combination of domain-specific financial statistics and signal features performs very well in this binary classification task. By ensembling a large number of leaf-wise growing trees, the proposed approach results in a noticeable increase of the forecasting accuracies, e.g., 8% for the real estate industry and nearly 6% for the cement and coal industries, respectively, as compared with the XGB model.

Table 7 also shows that the ARMA model performs better than the SVM model and approaches the accuracies of the proposed method on the real-estate and coal industries. However, the ARMA has a much worse performance when evaluated on the other two industries. The ARIMA model does not deliver good performance across these four industries and is not considered a suitable candidate for short-term predictions, which may be attributed to the fact that it

Table 7: Forecast accuracies (%) of the real estate, electric power, cement, and coal industries.

Industry	Real estate	Electric power	Cement	Coal
SVM	60	54	55.9	58.5
RF	63.8	57.3	50.8	51.5
SVM + RF	62.9	55.5	55.9	53.7
XGB	63.4	62.7	68.9	57.3
ARMA	68.3	50.9	61.9	62.3
ARIMA	32.7	49.1	40.0	14.7
Proposed	71.5	64.3	75.2	63.7

eliminates the influence of fluctuation trends by including a differential operation in the computation. It is worth mentioning that the proposed method does not access at all to the validation data and hence has better generalization capabilities. It effectively uses a combination of financial indicators and mechanic-specific signal features, which are obtained based on only the training data. On the contrary, we have to resort to the validation data in constructing the ARMA model in the short-term trend analysis.

4. Conclusion

In this paper, we proposed a novel method to perform price trend prediction based on the LGBM model and a variety of feature engineering techniques. The wavelet transform is used to filter high-frequency noise from time sequences, thus alleviating instabilities inherent in the financial data analysis. Furthermore, we proposed to derive multidimensional features as inputs to the model based on domain-specific technical indicators and the expertise on the mechanical signal analysis. The derived features enable the model to deliver significantly better performance as compared with statistical models and conventional machine-learning algorithms. The proposed model, however, still requires a computationally intensive optimization of LGBM hyperparameters. For future work, we will investigate the ensembling of tree-based models and CNN models. Transformer architectures that incorporate long-range attention mechanisms will also be studied for sequence-to-sequence prediction tasks.

Data Availability

The stock price data used to support the findings of this study are included within the article and cited as reference [31].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was funded by the Weihai Beiyang Electrical Group Co. Ltd., China, and the Shandong University, China.

References

[1] M. Z. Asghar, F. Rahman, F. M. Kundi, and S. Ahmad, "Development of stock market trend prediction system using multiple regression," *Computational & Mathematical Organization Theory*, vol. 25, no. 3, pp. 271–301, 2019.

[2] S. W. Lee and H. Y. Kim, "Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation," *Expert Systems with Applications*, vol. 161, p. 113704, 2020.

[3] L. S. Han and M. J. Nordin, "Integrated multiple linear regression-one rule classification model for the prediction of stock price trend," *Journal of Computer Science*, vol. 13, no. 9, pp. 422–429, 2017.

[4] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in *Proceedings of the 16th international conference on computer modelling and simulation*, pp. 106–112, IEEE, Cambridge, UK, March 2014.

[5] O. Y. Grachev, "Application of time series models (ARIMA, GARCH, and ARMA-GARCH) for stock market forecasting," *Dissertation, Northern Illinois University*, Dekalb, IL, USA, 2017, https://commons.lib.niu.edu/handle/10843/17833?show=full.

[6] C.-F. Huang, "A hybrid stock selection model using genetic algorithms and support vector regression," *Applied Soft Computing*, vol. 12, no. 2, pp. 807–818, 2012.

[7] M. Nabipour, P. Nayyeri, H. Jabani, and A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," *IEEE Access*, vol. 8, pp. 150199–150212, 2020.

[8] Y. Shi, W. Dai, and W. Long, "A new deep learning-based zero-inflated duration model for financial data irregularly spaced in time," *Frontiers in Physics*, vol. 9, p. 245, 2021.

[9] H. R. Kim, S. H. Hong, and H. Hong, "Machine learning based stock price fluctuation prediction models of KOSDAQ-listed companies using online news, macroeconomic indicators, financial market indicators," *Journal of Korea Multimedia Society*, vol. 24, no. 3, pp. 448–459, 2020.

[10] G. Ke, Q. Meng, T. W. Finley, and T. Wang, "LightGBM: a highly efficient gradient boosting decision tree," *Neural Information Processing Systems*, pp. 3149–3157, 2017.

[11] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, and F. S. H. Oueslati, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting," *Energy*, vol. 214, Article ID 118874, 2021.

[12] H. Jiajun, Y. Chuanjin, L. Yongle, and X. Huoyue, "Ultra-short term wind prediction with wavelet transform, deep belief network and ensemble learning," *Energy Conversion and Management*, vol. 205, no. 1, Article ID 112418, 2020.

[13] A. Prochazka, J. Kukal, and O. Vysata, "Wavelet transform use for feature extraction and EEG signal segments classification," in *Proceedings of the The 3rd International Symposium on Communications, Control and Signal Processing*, March 2008.

[14] C. Dimoulas, G. Kalliris, G. Papanikolaou, and A. Kalampakas, "Long-term signal detection, segmentation and summarization using wavelets and fractal dimension: a bioacoustics application in gastrointestinal-motility monitoring," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 438–462, 2007.

[15] P. Johankhani, V. Kodogiannis, and K. Revett, "EEG signal classification using wavelet feature extraction and neural networks," in *Proceedings of the IEEE John Vincent Atanasoff International Symposium on Modern Computing*, IEEE, Sofia, Bulgaria, October 2006.

[16] D. Zhang, *Wavelet Transform. In Fundamentals of Image Data Mining*, Springer, New York, NY, USA, 2009.

[17] C.-L. Lin, *A Tutorial of the Wavelet Transform*, NTUEE, Taiwan, China, 2010.

[18] J. Guerrero-Turrubiates, S. Ledesma, S. Gonzalez-Reyna, G. Avina-Cervantes, and E. Ilunga-Mbuyamba, "Guitar audio signal classification by collapsed pitch class profile," in *Proceedings of the IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, November 2016.

[19] X. Guo, "Chapter 11 clustering of NASDAQ stocks based on elbow method and K-means," *Springer Science and Business Media LLC*, vol. 2021, 2021.

[20] K. Raza, "Prediction of Stock Market performance by using machine learning techniques," in *Proceedings of the International Conference on Innovations in Electrical Engineering and Computational Technologies*, Beach Resort Phuket, Thailand, October 2017.

[21] B. Ding and L. Li, "Research on comprehensive analysis method of stock KDJ index based on K-means clustering," in *Proceedings of the 3rd ICMEIT*, Dalian, China, March 2019.

[22] Y. Shi, B. Li, W. Long, and W. Dai, "Method for improving the performance of technical analysis indicators by neural network models," *Computational Economics*, vol. 59, no. 3, pp. 1027–1068, 2021.

[23] Y. Qu, Z. Zhang, and Z. Qin, "Wavelet-Aided stock forecasting model based on ensembled machine learning," in *Proceedings of the The 3rd International Conference on Machine Learning and Machine Intelligence (MLMI)*, Hangzhou, China, September 2020.

[24] J. J. Welles Wilder, *New Concepts in Technical Trading Systems*, Trend Research, Kingston, NY, USA, 1978.

[25] R. Jegadeeshwaran and V. Sugumaran, "Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines," *Mechanical Systems and Signal Processing*, vol. 52-53, pp. 436–446, 2015.

[26] R. S. Figliola and D. Beasley, *Theory and Design for Mechanical Measurements*, John Wiley & Son, Hoboken, NJ, USA, 2015.

[27] M. Xia, T. Li, L. Xu, L. Liu, and C. W. Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE*, vol. 23, no. 1, pp. 101–110, 2018.

[28] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, pp. 522–542, 2021.

[29] D. Jia and R. Xue, "Research on earnings management of growth enterprise market in China Stock Market: comparative analysis based on the BPNN, GBDT, and MLR models," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6064536, 2022.

[30] J. Yu, Q. Lu, Q. Lu et al., "A multi-stage ensembled-learning approach for signal classification based on deep CNN and LGBM models," *Journal of Communications*, vol. 17, pp. 30–38, 2022.

[31] TuShare 043, "TuShare is a utility for crawling historical data of China stocks," 2017, http://tushare.org/index.html.

[32] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *Applied Mathematical Finance*, vol. 27, 2020.

[33] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the The 22nd ACM international conference on knowledge discovery and data mining*, pp. 785–794, San Francisco, CA, USA, August 2016.