IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON VISUAL ANALYSIS FOR CPS DATA

Received April 30, 2019, accepted May 15, 2019, date of publication May 27, 2019, date of current version June 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919189

Novel Financial Capital Flow Forecast Framework Using Time Series Theory and Deep Learning: A Case Study Analysis of Yu'e Bao Transaction Data

XIAOXIAN YANG[1], SHUNYI MAO[2], HONGHAO GAO[2], YUCONG DUAN[3], AND QIMING ZOU[2,4]

[1]School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China
[2]Computing Center, Shanghai University, Shanghai 200444, China
[3]College of Information Science and Technology, Hainan University, Haikou 570100, China
[4]Shanghai Shangda Hairun Information System Co., Ltd., Shanghai 200444, China

Corresponding author: Shunyi Mao (specialyj@shu.edu.cn)

This work was supported by the Youth Foundation of Shanghai Polytechnic University under Grant EGD18XQD01, in part by the CERNET Innovation Project under Grant NGII20170513, and in part by the National Natural Science Foundation of China under Grant 61502294.

ABSTRACT Appropriate monetary liquidity is important for financial institutions. When institutions lack adequate cash flow for customer redemption, their income will decrease, their reputation will be affected, and they may even go bankrupt. However, the opposite extreme in which more cash is reserved than needed may result in lost opportunities to make successful investments. This study uses Yu'e Bao transaction data to investigate a method for forecasting financial capital flow. Yu'e Bao, which is a financial product launched by Alibaba, faces the core challenge of maximizing commercial profits to reduce investment risks. Liquidity risk is considered the main factor in Yu'e Bao's investment strategy. First, a linear model called YEB_ARIMA is proposed by determining the autocorrelation (ACF) and partial autocorrelation (PACF) parameters, which are optimized by the grid search method. Second, a deep learning model called YEB_LSTM is introduced to strengthen the expressiveness of the model that yields nonlinear transaction features. Then, a hybrid learning method called YEB_Hybrid is applied to improve the original weak classifiers. This model includes both a linear combination and logistic regression learning. Third, a set of experiments and analyses are conducted based on subscription and redemption datasets to demonstrate that the hybrid model achieves an accuracy of 84.39% and 84.36%, respectively, under a variety of evaluation indexes. Finally, various proposed fund reserve ratios are provided based on capital forecasts.

INDEX TERMS Liquidity risk, ARIMA, LSTM, time series, capital flow prediction, big data financial analysis.

I. INTRODUCTION

Financial markets have changed dramatically due to the development of the Internet, new financial products, mobile financial capabilities [1], and risk control systems [2]. These changes have introduced enormous volatility and uncertainty to the financial world, although they also indicate that the time for a sea change in the status of finance is imminent [3]. In June 2013, a sudden ''money shortage'' hit China. The imbalanced money demand and supply caused the interbank rates to soar [4]. During this period, the financial product Yu'e Bao, a joint venture of Alibaba and Tianhong Asset Management, found an opportunity. Through 2016/6, Yu'e Bao's customer base reached 2.95 hundred million users, which easily outpaced the original market leader. This growth occurred because Yu'e Bao effectively avoids high thresholds for participating in finance as well as tedious registration procedures. Yu'e Bao's ultralow one-yuan financial threshold enables people with even small amounts of idle funds, such as students, to participate in financial markets, thus opening up an era of national financial management.

Financial forecasting plays an important role in addressing uncertainty regarding the future and analyzing variation

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

trends. Traditional capital forecasts mainly adopt qualitative theories whose subjectivity is strong and generalization ability is weak. The traditional method for performing quantitative predictions [5] uses a relatively simple linear model. Although the scientific basis of this model has been improved to a certain extent, its predictive accuracy is poor and theoretical foundations are relatively weak. Modern research methods basically combine qualitative and quantitative methods. In July 2017, the state promulgated the policy captured in a white paper on the development plan for a new generation of artificial intelligence [6], which provides a powerful tool for financial liquidity risk prevention and control [7]. Time series theory analyzes data by adopting statistical methods and extracts meaningful insights about them. Long short-term memory (LSTM) networks are better at considering the information in long sequences than traditional RNN networks, although they are both effective tools [8] for making financial forecasts. Through theoretical analysis and data modeling, this study predicted the future capital flow of Yu'e Bao with the goal of reducing the company's management risk and maximizing its benefits.

For Yu'e Bao, the sudden growth of users and capital flow are both a rare opportunity and a large risk. In the face of huge daily volumes, ensuring future cash flow is crucial. The biggest risk faced by Yu'e Bao is closely related to its customers, namely, fund liquidity [9]. Liquidity risks arise mainly from banks' inability to cope with liquidity difficulties caused by falling liabilities or rising assets. When a financial institution lacks liquidity, it must sell investment products early to increase capital, thus affecting its profitability. In extreme cases, illiquidity can cause financial institutions to go bankrupt. Therefore, analyzing users' behavioral data when interacting with Yu'e Bao, summarizing their characteristics, and exploring a more reasonable method to forecast cash flow are of great importance not only to Yu'e Bao's operation but also to a wide range of enterprises.

The remainder of this paper is organized as follows. Section 2 presents the motivation and methods of building a capital flow prediction framework and introduces the main characteristics of Yu'e Bao's dataset. Section 3 describes the process of building YEB_ARIMA, including determining the difference order and the autocorrelation (ACF) and partial autocorrelation (PACF) orders. Section 4 describes the construction of YEB_LSTM while considering correlations between time series, including feature selection, network design, etc. Section 5 presents the ensemble methods that combine YEB_ARIMA and YEB_LSTM to improve the prediction accuracy and then calculates the reserve ratio. Section 6 evaluates the models from three aspects: residual noise, Ljung–Box Q (LBQ), and mean average prediction error (MAPE). Section 7 describes related works. Finally, Section 8 presents the conclusions and future work.

II. MOTIVATION

Yu'e Bao is an open-ended money fund that can be traded at any time. Investors can purchase or redeem Yu'e Bao products at any time according to their own financial needs. Figure 1 shows the importance of fund flow statistics for companies. When users' redemptions are greater than the company's reserve funds, a liquidity risk occurs that will reduce the company's performance and negatively impact investors. In the most serious case, the company may face a threat of bankruptcy. In contrast, when users' redemptions are less than the company's reserve funds, a liquidity surplus is generated. This surplus means that the company is not able to make full use of the money at hand and cannot maximize benefits, which can be considered a loss to both the company and its investors.

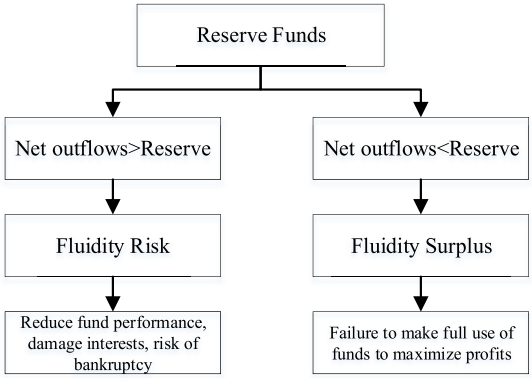

FIGURE 1. Importance of cash flow statistics.

Big data risk control has become an important application under the themes of "artificial intelligence (AI)" and "big data" [10]. Internationally renowned Internet giants, such as Google, IBM, Alibaba, Baidu and Tencent, have applied AI to analyze various financial services. Well-known financial industry names, such as JPMorgan Chase, Goldman Sachs, and accounting firms, such as PWC and Ernst & Young [11], have all followed this technological wave, and many have invested heavily in the AI field.

The complete dataset of Yu'e Bao is commercially protected. Therefore, the model established in this paper adopts a smaller dataset published by Yu'e Bao, namely, the Purchase_Redemption_2010 dataset. This dataset contains 2,800,000 transaction records of 30,000 Yu'e Bao users for the period from July 2013 to August 2014. To illustrate the daily trading situation, the daily fund subscription amount and redemption amount are visualized in Figure 2. It can be seen and validated that when the purchase amount arises, the redemption amount goes up as well; thus, they have a positive correlation. From July 2013 to October 2013, the transaction amount was small since Yu'e Bao was not popular when first entering the market. Then, its transaction amount gradually increased and varied greatly, which might result in very large fluidity risks. Figure 3 shows the process of modeling Yu'e Bao's capital flow. Yu'e Bao 's subscription and redemption funds, annualized interest rate, and user consumption are used as the input data. The YEB_ARIMA model was established based on these input

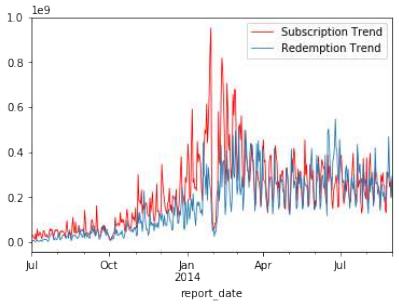

FIGURE 2. Total subscription and redemption amounts.

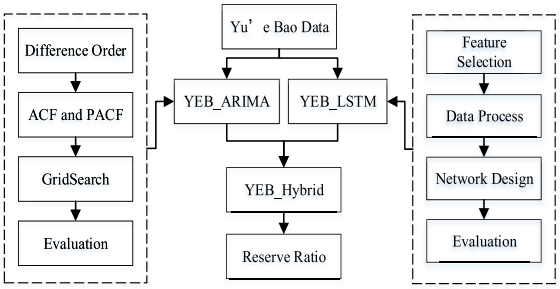

FIGURE 3. Yu'e Bao capital flow model process.

data, and the data difference order, ACF coefficient and PACF coefficient were observed. Finally, the YEB_ARIMA (p,d,q) was determined and the predicted values of Yu'e Bao fund purchases and redemptions were obtained. Then, the YEB_LSTM short-term and long-term memory neural network model was established based on the input data. To create this model, the key features useful in modeling had to be identified and the raw data had to be preprocessed to ensure that they were suitable for neural network characteristics. Finally, the entire network structure was designed and the results of model predictions were obtained.

After acquiring the predicted results of the linear model YEB_ARIMA and the nonlinear model YEB_LSTM, the integrated learning stacking method was adopted. In this approach, the goal is to fuse the two models using certain strategies to improve the overall prediction effect. Based on the daily purchase and redemption results predicted by this fusion model, the size of the daily reserve ratio was determined, which can help reduce the risks of liquidity shortfalls and surpluses. This knowledge could help the company improve its capital flow management, avoid risks, and maximize benefits.

III. MODEL 1: TEMPORAL RELATIONS USING YEB_ARIMA

In time series prediction, linear and nonlinear methods have been studied [12] by many scholars. The linear models mainly include autoregression (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and other models, while the nonlinear models mainly include the hidden Markov chain (HMM) and the artificial neural network (ANN). To choose a suitable model for the time series, both the characteristics of the time series and the characteristics of the model should be analyzed first. Three factors, i.e., the short-term autocorrelation of sequences, the selection of a linear model and the stationarity of sequences, have important effects on time series modeling. Figure 4 shows the process of selecting a linear model [13], and Figure 5 illustrates the modeling process for YEB_ARIMA. After the difference operation, the stationarity of the time series is tested and the difference order d is determined. Then, based on the sequence, the optimal autocorrelation coefficient p and the partial correlation coefficient q [14] are selected.

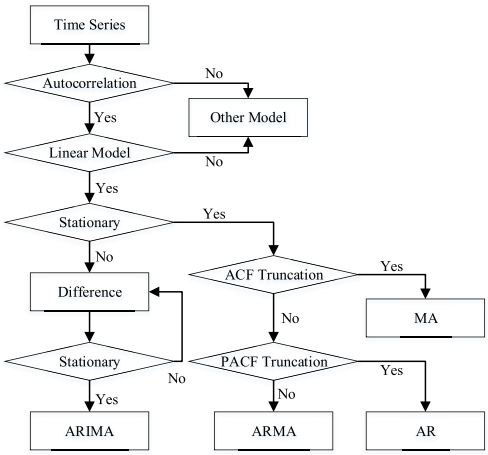

FIGURE 4. Linear model selection process.

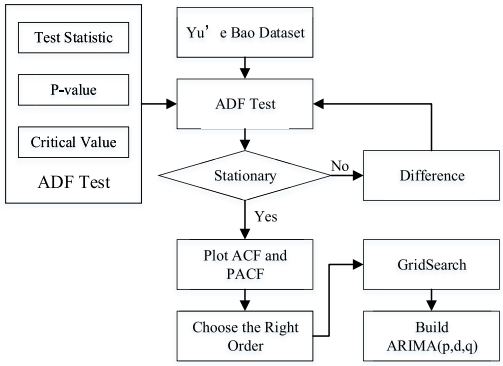

FIGURE 5. YEB_ARIMA modeling process.

A. DETERMINE THE DIFFERENCE ORDER

According to the user sequence scatter plot, the stationarity of the sequences must be analyzed during the investigation sequence when steady properties exist. If there is a seasonal trend, the seasonality should first be eliminated by applying a sliding window averaging operation; then, the sequence should be judged for stability by the asymptotically distribution-free (ADF) test. If the time series is

TABLE 1. ADF checklist of Yu'e bao data.

	Subscription		Redemption	
Data type	Origin	First Difference	Origin	First Difference
Test Statistic	-1.59	-7.95	-1.37	-11.69
p-value	0.49	3.19e-12	0.59	1.64e-21
Number Used	408.00	407.00	413.00	413.00
Critical Value (1%)	-3.45	-3.45	-3.45	-3.45
Critical Value (5%)	-2.87	-2.87	-2.87	-2.87
Critical Value (10%)	-2.57	-2.57	-2.57	-2.57

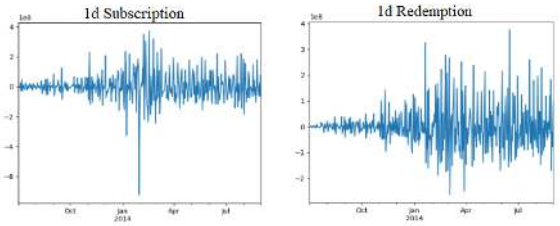

FIGURE 6. First-order difference.

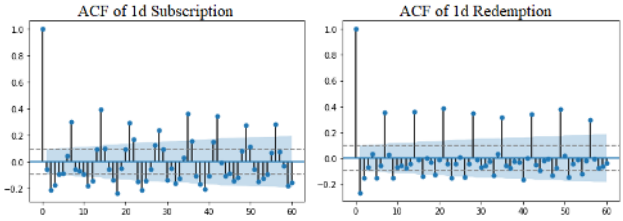

FIGURE 7. ACF of first-order difference data.

a nonstationary state, then differential and other operations must be performed on the original sequence until the state as tested by ADF is stable.

The stability of the time series is investigated according to the ADF test values [15] presented in Table 1. The main analysis process is as follows:

1. If the test statistic is less than the critical value, the sequence may be stable. The 1%, 5% and 10% designations after the critical value represent the significance levels of the tests. A smaller critical value indicates a more stable sequence.
2. The closer the p-value is to 0, the better its ability to verify that the sequence is stable.

The ADF test table of all the subscription data reveals that the test statistic of the original data is -1.59 while the critical value of the 10% significance test is -2.57. Thus, the size of the test statistic is larger than the critical value (10%); therefore, the sequence is not stationary. In addition, the p-value is 0.488, which is far from 0; consequently, the original redemption data sequence is not stable. To improve the stability of the sequence, a differential operation was applied to the original time sequence. After the first-order difference operation for the subscription sequence, the test statistic was -7.947 and the critical value was -3.446 when the significance level was 1%. This test value is less than 1%; therefore, the sequence is stable. In addition, the size of the p-value is 3.198e-12, which is very close to 0. Therefore, all the redemption data after the first-order difference operation are considered to be stable. By comparing the ADF values after the first-order difference for the redemption data, we found that the first-order difference operation is also a stable sequence. Thus, the first-order difference operation is selected for the subscription data and redemption data.

B. DETERMINE THE ACF AND PACF

When the processed time series data are stationary, a linear time series model can be constructed. ACF and PACF diagrams of the sequence are drawn to observe the autocorrelation and partial autocorrelation coefficient of the sequence. The ACF [16] reflects the value of a time series at time τ and the degree of linear dependence between time series at time $\tau + h$. The PACF measures the amount of correlation between a variable and the lag of itself, which ignores the correlations at all lower-order lags. The AR(p) model is established when the ACF diagram observed in the sequence shows a trailing phenomenon, while the PACF diagram shows a truncated phenomenon. In contrast, the MA(q) model is established when the ACF diagram is truncated and the PACF coefficient exhibits a tail. When both the ACF and PACF diagrams show a trailing phenomenon, it should be considered whether the time series has been transformed into a stationary sequence after the differential operation. If it is stabilized by the difference, an ARIMA(p,d,q) model is established; otherwise, an ARMA(p,q) model is established [13].

Figures 7 and 8 show that the order of the model can be roughly obtained. First, the autocorrelation coefficient of the first-order difference data was observed. Strong autocorrelations exist between the subscription data and redemption data every 7 days. By observing the partial autocorrelation coefficient of the first-order difference data, it can be found

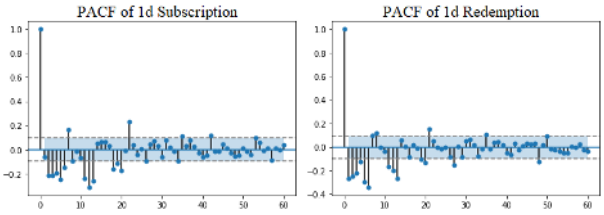

FIGURE 8. PACF of first-order difference data.

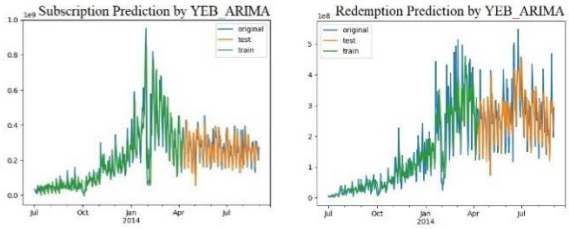

FIGURE 9. YEB_ARIMA modeling result.

that the partial autocorrelation coefficient of subscription and redemption has a tail truncation trend after the 12th order. However, in the actual modeling process, the observed order is often not able to achieve the optimal modeling effect; then, the optimal model parameters must be found via optimization.

According to the YEB_ARIMA modeling process, the grid search method [17] can be used to optimize the parameters of the subscription and redemption data to obtain the optimal model parameters. In this study, the orders of the subscription and redemption data were determined using the grid search method. The final subscription model parameters were YEB_ARIMA(8,1,7), and the redemption model parameters were YEB_ARIMA(1,1,8).

C. YEB_ARIMA PREDICTION RESULT

After determining how to select the model and the model order, experiments were conducted using the Yu'e Bao dataset. The experiment environment for this study is a Lenovo (Beijing) laptop computer running Windows 10 with 8 GB of memory and an Intel Core i5-6200U CPU @2.3 GHZ. The software environment included the Pycharm editor and Python 3.5, which were installed by Anaconda.

YEB_ARIMA uses the first 2/3 of the data as the training set and the last 1/3 data as the testing set. Figure 9 shows the results of the subscription model YEB_ARIMA(8,1,7) and the redemption model YEB_ARIMA(1,1,8). The blue portion represents the original data, the green portion represents the predicted data for the training set, and the orange portion represents the test data predictions. Note that the predicted values start several days later compared with the original data because modeling through YEB_ARIMA uses the data from the previous cycle as the basis for predicting the funds required for the next day.

$$MAPE = \sum_{i=1}^{n} \left| \frac{observed_i - predicted_i}{observed_i} \right| \times \frac{100}{n} \quad (1)$$

MAPE is used to assess the regression error [18] between the proposed model result and the ground truth. The mean squared error (MSE) and root mean square error (RMSE) are additional regression evaluation indexes; however, they are more easily influenced by abnormal values than MAPE. YEB_ARIMA focuses on the timing characteristics of the Yu'e Bao data, including their stationarity, autocorrelation and partial autocorrelation, to model the future capital flow, and it achieved an accuracy of 82.62% on the subscription data and an accuracy of 80.99% on the redemption data.

IV. MODEL 2: NETWORK FEATURE-BASED YEB_LSTM

To establish the neural network, in addition to considering the correlations between time series, more attention should be paid to the network structure design. For example, how to perform data cleaning to meet the network requirements and how to choose the most appropriate features for the network modeling are both vital tasks. In this study, the time series predictions of Yu'e Bao subscription and redemption data must estimate the users' future purchase and redemption amounts at time $\tau + T$ from the capital flow at time τ.

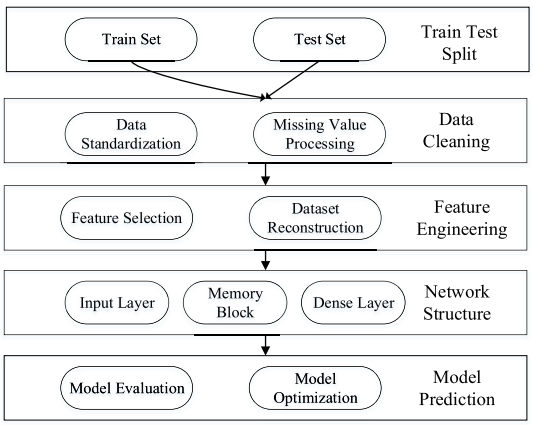

FIGURE 10. YEB_LSTM modeling process.

Figure 10 illustrates the complete modeling process for the YEB_LSTM model. The dataset is first split into a training set and a test set to validate the model effect. Second, a preprocessing step is performed, including data cleansing and data standardization. Then, the feature engineering process, which includes feature selection and dataset reconstruction, is conducted to meet the network input demands. Then, the input layer, memory block and dense layer of the YEB_LSTM network are designed. Finally, model evaluation and optimization are performed. These last three steps are briefly introduced below.

A. CRUCIAL FEATURE SELECTION

Before designing the network structure, the crucial features should be selected. We assume that the cycle of Yu'e Bao is T.

Note that when predicting the subscription and redemption amount of Yu'e Bao at time τ, it is necessary to use the abovementioned feature of Yu'e Bao at time τ-T.

First, the interbank interest rate is chosen. When a bank is short of capital, the interbank lending rate will increase. Consequently, the agreed deposit rate of Yu'e Bao will increase and the annualized interest rate return to users will also increase.

Second, the annualized interest is selected. The annualized interest rate of Yu'e Bao shows that when the annualized interest rate is relatively high (e.g., around January and February 2014), a large number of fund applications occur. Similarly, when the annualized interest rate is low, a large number of redemptions occur.

Third, investor consumption is also vital. User consumption over several special holidays is enormous, such as the ''double 11'' day promoted by e-commerce. On November 11, a total of 16.79 million redemptions were made from Yu'e Bao: the amount reached 6.125 billion yuan, which represents the largest single-day redemption record in the fund's history [19]. Consumption situations have a great influence on redemption quantity.

Fourth, fund transfer is used. The subscription volume of Yu'e Bao is relatively high at the beginning of the month because shortly after the launch of Yu'e Bao, most banks put a month limit on the amount of money that can be transferred to Yu'e Bao. Therefore, when the monthly quota of users occurs at the beginning of each month, there will be a large number of applications.

Fifth, subscription and redemption data are needed. The analysis in the YEB_ARIMA model includes subscription and redemption data autocorrelations and partial autocorrelations. Furthermore, the Yu'e Bao data are periodic.

B. NETWORK STRUCTURE DESIGN

1) YEB_LSTM STRUCTURE

Several network structures have been applied to analyze financial time series; however, they have some defects. In traditional dense neural networks [20], the output of the preceding neuron layer and the input of the subsequent neuron layer are usually fully connected; however, there is no correlation between the inner neuron nodes in each layer of the network, which ignores any mutual influences of the data. In addition, a recurrent neural network [21] is used for the series analysis, and it faces the vanishing gradient problem. The LSTM [22], which uses three types of gated memory units, effectively avoids the vanishing gradient problem that can occur during backpropagation. Moreover, a LSTM can remember the inherent trends in the series data; thus, it is suitable for financial data analysis. LSTM adds a series of mechanisms to the traditional circular neural network. These specific methods are as follows.

A forgetting mechanism was added first. For example, the subscription amount of Yu'e Bao at time τ may be related to its value at time τ-T but not related to the amount at the previous time. Here, T is selected to be 7. Therefore, when training the network model, we hope to eliminate the effect at time τ-8, τ-9 and so on. By adding this forgetting mechanism, the LSTM learns what information can be forgotten when new data enter the network. This mechanism also serves to preserve the most effective historical information. When the LSTM model obtains a new subscription or redemption value, the model needs to know which unused long-term information [23] should be forgotten and then store the useful information in its long-term memory. The model maintains the long-term memory in a separate working route. When the model is trained, it knows which part of the information can be used directly; that is, it knows the key information stored in long-term memory.

The architecture of the YEB_LSTM model designed in this paper is shown in Figure 11. The specific structure of this network is as follows: 1) Five sets of features are input, and each feature set may contain one or more specific features for a total of 12 input nodes; and 2) the first layer is the LSTM layer, which contains 10 hidden nodes. The YEB_LSTM has a timestep of 7 in this paper since the dataset is constructed to use the past 7 days' information to predict the current day's amount. At every timestep, the YEB_LSTM has an output. However, only the last output of the YEB_LSTM is chosen to be the input of the next dense layer. The second layer is the dense layer, which contains 8 nodes; 3) The last layer is the output layer, which yields the forecast amount for Yu'e Bao subscriptions or redemptions.

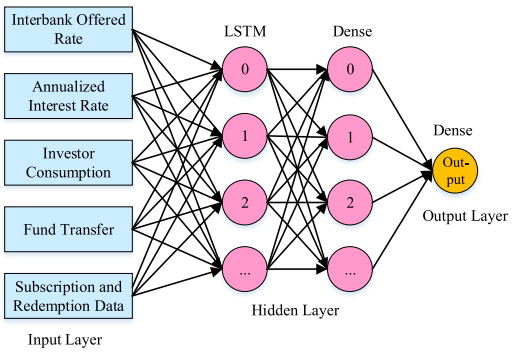

FIGURE 11. Structure of YEB_LSTM.

2) YEB_LSTM DETAILS

a: SUPERPARAMETER SELECTION

In the YEB_LSTM network described in this paper, the superparameters are set as follows: an epoch equals 100, which means the number of learning iterations during the training process is 100. The learning rate is 0.05, which is the step size when solving the gradient descent. The small batch size equals 1, which means that the training of small batch gradient descent is conducted each time a data item is input.

b: LOSS FUNCTION

To measure the phenomenon of excessive or insufficient fitting to the data, the role of the loss function [24] is to measure

the gap between the predicted value and the real value. In this experiment, the mean squared error (MSE) loss is used during training.

c: OPTIMIZATION METHOD

To reduce the error measured by the loss function, an optimization method [25] should be adopted to improve the original model parameters. The optimization method adopted in this study is the Adam optimization method.

d: ACTIVATION FUNCTION

When using a deep learning method to train the model, an activation function is often added to increase the nonlinear expression ability of neural networks. In this paper, sigmoid activation functions are used:

$$sigmoid(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

C. YEB_LSTM PREDICTION RESULT

YEB_LSTM uses the first 2/3 of the data as the training set and the last 1/3 of the data as the testing set. The experimental environment and the error metric MAPE are the same as those used for the YEB_ARIMA model experiments. Figure 12 shows the results of the YEB_LSTM model. The model achieved accuracy rates of 79.41% and 81.25% on the Yu'e Bao subscription and redemption datasets, respectively.

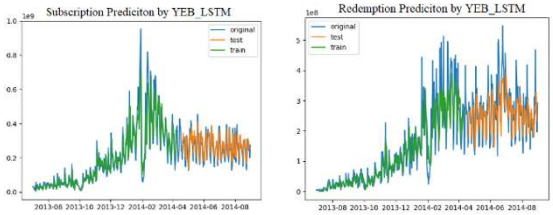

FIGURE 12. YEB_LSTM modeling results.

V. ENSEMBLE MODEL YEB_HYBRID AND RESERVE RATIO

Single models are fast and relatively easy to build; however, their accuracy is usually lower than that of ensemble models [26]. One effective method of improving model accuracy is to train multiple independent models and then combine them using various strategies. This approach is called ensemble learning [27]. As shown in Figure 13, the YEB_Hybrid model uses the ensemble learning method by combining different classifiers to improve the system's predictive power, which often achieves significantly better generalization performance than single learners. In this paper, the YEB_ARIMA model for sequential relationships is a linear model with relatively complex modeling steps, although its prediction process is relatively fast, while the network-oriented YEB_LSTM is a nonlinear model that requires a relatively long training process but achieves good results. Thus, it is desirable to combine the linear

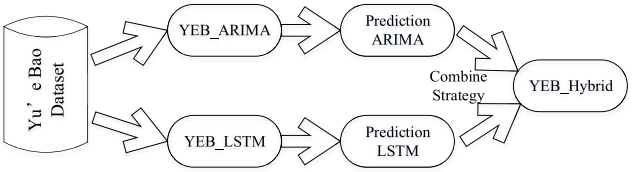

FIGURE 13. YEB_Hybrid model process.

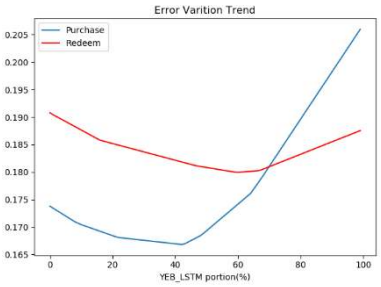

FIGURE 14. Linear fusion trend chart.

YEB_ARIMA model and the nonlinear YEB_LSTM model because it can help to observe data from different perspectives. After understanding the principle of integrated learning, the original results of the classification experiments must be processed to obtain the results of the integrated learning model. As Figure 14 shows, the prediction error could be optimized via linear combination.

A. ENSEMBLE METHOD

Table 2 summarizes the results of modeling with the linear combination method. As the table shows, the model error effect after using the linear combination method is better than that of the independent model. This finding shows that it is feasible to optimize the model through an integrated learning approach. Table 3 shows that the prediction accuracy of the YEB_ARIMA and YEB_LSTM models is relatively high and the error is held to approximately 20%. By using the linear combination method, the prediction effect is improved to some extent after assigning different weights to the basic models. Using the regression learning method in the integrated learning approach is even more effective, and it reduces the error rate of subscription to 15.61% and the error rate of redemption to 15.64%. Therefore, for the YEB_Hybrid model, the result of the logical regression combination is taken as the final result.

B. FUND RESERVE RATIO

In Yu'e Bao's fund operation, the size of the fund reservation ratio is easier to understand. In this section, we calculate the required daily fund reserve proportion based on the predicted fund subscription and redemption flows. The fund reserve ratio [28] is calculated as follows:

$$\xi = 1 - \phi - \varphi - \frac{S_\tau + P_\tau(Pur_\tau - Re_\tau)}{F_\tau} \tag{3}$$

TABLE 2. Linear combination modeling errors.

Model	Subscription			Redemption		
	B.F. (%)	Portion	A.F. (%)	B.F. (%)	Portion	A.F. (%)
YEB_ARIMA	17.38	0.58	16.68	19.08	0.39	17. 99
YEB_LSTM	20.59	0.42		18.75	0.61	

Explanation: B.F. = before fusion; A.F. = after fusion

TABLE 3. Prediction errors of different models.

Model		Subscription	Redemption
Basic Classifier	YEB_ARIMA	17.38%	19.08%
	YEB_LSTM	20.59%	18.75%
Ensemble	LinearCombination	16.68%	17.99%
	LogisticCombination	15.61%	15.64%

where all the parameters in the equation are included in the learning process at moment τ; F_τ represents the total amount of funds in Yu'e Bao; S_τ denotes the total number of shares sold; ξ is the calculated fund reserve ratio; ϕ represents company overhead, including the fund custody fee, staff service fee, etc.; φ is the proportion of a company's investments in stocks and other financial products; P_τ represents the unit price of Yu'e Bao at the current moment; and Pur_τ and Re_τ are the investor subscription and redemption ratios, respectively.

Due to the limited financial data released by Yu'e Bao, up to the first quarter of 2015, Yu'e Bao had 7.12×10^{11} yuan of funds. To calculate the funds for a certain day, the total can be assumed to be the daily average for the quarter. For example, these data would be adopted as the average daily fund amount of Yu'e Bao for August 2014. Yu'e Bao charges a 0.08% hosting fee, the management fee for the company's personnel is 0.3%, and the after-sales service and other expenses are 0.25%, for a total of 0.63%. Approximately 85 percent of the fund is invested in securities. We substitute these data into the original formula and obtain the following:

$$\xi = 1 - 0.63\% - 0.85 - \frac{P_\tau(Re_\tau - Pur_\tau) - 4.8 \times 10^{10} \div 90}{7.12 \times 10^{11} \div 90} \tag{4}$$

According to the above formula, for Yu'e Bao from July 27, 2014 to July 31, 2014, it is necessary to first calculate the subscription and redemption fund amounts derived from the YEB_Hybrid model. Table 4 lists the prediction results, where Subs_amt is the predicted subscription amount and Redempt_amt is the predicted redemption amount. Then, the predicted subscription and redemption amounts are used to calculate the capital reserve ratio predicted portion (Pred_por). The actual portion (Actual_por) is calculated using the actual data, and Error represents the prediction deviation.

TABLE 4. Yu'e Bao fund reserve ratio forecast.

Date	Subs_amt	Redmpt_amt	Predict_por (%)	Actual_por (%)	Error (%)
7/27	1.34e8	2.47e8	9.06	7.80	15.83
7/28	2.78e8	3.33e8	8.32	7.30	13.97
7/29	3.12e8	3.02e8	7.49	8.58	12.66
7/30	2.66e8	2.93e8	7.97	8.14	2
7/31	2.49e8	2.49e8	7.62	8.71	12.46
Average Prediction Error					11.38

Explanation: Subs_amt = subscription amount; Redmpt_amt = redemption amount

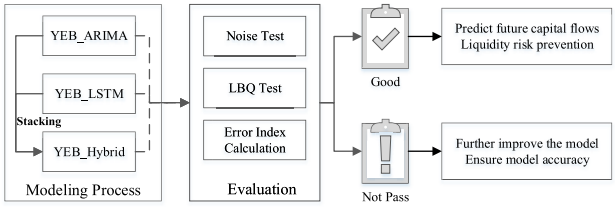

FIGURE 15. Model evaluation process.

VI. MODEL EVALUATION

In addition to minimizing the deviation between the predicted value and the real value, a good model should also make full use of the information in the input data. The model evaluation method adopted in this paper is mainly divided into three parts: the residual noise characteristic test, the LBQ test and error index calculation. All three models are checked using these methods; and we take the YEB_Hybrid model as an example.

A. RESIDUAL NOISE CHARACTERISTIC TEST

In the model evaluation, the residual between the ideal modeling result and the real data [29] should consist of white noise, which means that no correlations occur between variables, the mean value is 0, and the residual satisfies a Gaussian distribution.

The residual error between the predicted results of the YEB_Hybrid model and the real data can be intuitively seen from Figure 16, where the sequence is random and changes

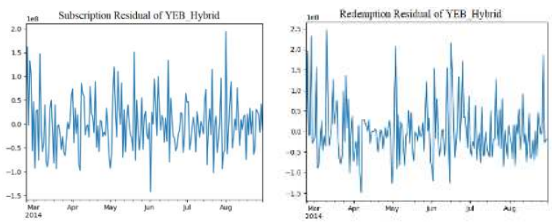

FIGURE 16. YEB_Hybrid residual.

TABLE 5. YEB_Hybrid residual ADF test.

	Subscription	Redemption
Test Statistic	-15.46	-13.91
p-value	2.71e-28	5.64e-26
Sample Counts	188.00	188.00
Critical Value (1%)	-3.46	-2.58
Critical Value (5%)	-2.58	-2.88
Critical Value (10%)	-2.88	-3.47

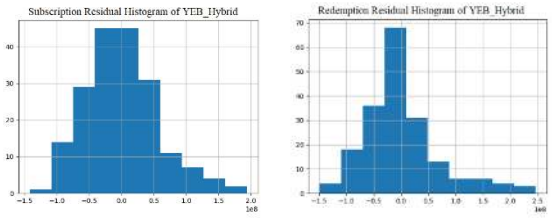

FIGURE 17. YEB_Hybrid residual histograms.

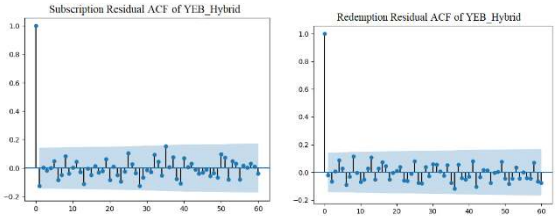

FIGURE 18. YEB_Hybrid residual ACF test.

irregularly around the 0 mean. To objectively evaluate the YEB_Hybrid modeling results, the ADF test is conducted on its residual to observe whether the sequence is stable.

As shown in the noise histogram in Figure 17, the peak of the residual data is located near the 0 position and the left and right sides are roughly symmetrical, which confirms the requirement that Gaussian white noise [30] should have a normal distribution. In Figure 18, the autocorrelation diagram shows that the correlation test values are all within 5%, indicating that no significant correlation exists in the residual sequence. This result conforms to the requirement that Gaussian white noise should have no autocorrelation.

B. LBQ TEST

The LBQ test is mainly used to investigate whether a lag correlation exists between data. When the significance value is smaller than 0.1, the model is not suitable. Otherwise, a significance value closer to 1 [31] indicates a better-fitting model. Table 6 shows that the LBQ test values of the YEB_Hybrid model is all larger than 0.1. This result means that YEB_Hybrid makes full use of the effective information in the data and that no autocorrelation exists between the residuals. This result also conforms to the definition that no autocorrelation should exist in Gaussian white noise.

TABLE 6. YEB_Hybrid residual LBQ test.

Lag	1	2	3	4	5	6
Subs.	0.81	0.97	0.99	0.99	0.99	0.99
Redem.	0.97	0.91	0.97	0.64	0.76	0.70

Explanation: Subs = Subscription; Redem = redemption

TABLE 7. Predicted results of the reference.

Group	Prediction Error
Inactive	33.49%
Active	18.30%
Mean Error	25.90%

C. MAPE ERROR CALCULATION

The subject of this research stems from the Yu'e Bao fund flow prediction competition held by the Tianchi big data platform on 2015/8/7. Table 7 shows the optimal forecast results given by the [32], which divide users into active and inactive groups that have errors of 33.49% and 18.30%, respectively.

In this paper, the subscription prediction error was 15.61%. On the redemption dataset, the predicted error was 15.64%, which is far lower than the reference result. This comparison with the reference experiments shows that the modeling in this paper greatly improves the accuracy of prediction and achieves a good effect. Consequently, the ensemble model YEB_Hybrid achieves 84.39% and 84.36% accuracy for the subscription and redemption processes, respectively, and performs well under multiple evaluation indexes.

VII. RELATED WORKS

Many scholars have performed relevant work on the risk factors and risk control of the fund market. IEEE CIFEr [33] reports that many large banks and financial companies have made massive investments to improve risk assessment and control by using mathematics, computer technology and financial engineering knowledge, such as machine learning methods. Agarwal *et al.* [34] found that funding liquidity and leverage ratio are the main determinants of hedge fund performance, and a tail risk measure was developed. Yongxin [35] noted in the research that as the number and types of fund companies become increasingly large, they are more likely to face redemption risks in a competitive environment. Future fund returns affect user behavior; therefore, it is extremely important to make advance fund predictions to prevent liquidity risks. Sun *et al.* [36] introduced a portfolio optimization method that used gradient descent to maximize the cash flow when pension fund terms expire. Zhang *et al.* [37] used dynamic programming to formulate a task to minimize insurance risk and maximize profit into a stochastic differential game. Yan and Ouyang [38] combined a LSTM deep learning model with and wavelet analysis to capture the complex

features of financial time series and verify the effectiveness of LSTM in making time series predictions. Yang *et al.* [39] used a deep learning method to perform automatic feature engineering and predict whether traders will secure profits. Liu *et al.* [40] adopted a deep learning network called ResNet50 to process financial data, and the bagging method is also used to perform the classification. Ahmed *et al.* [41] utilized machine learning methods, such as clustering and nearest neighbors, to detect anomalies in financial data. These studies indicate that combining machine learning, deep learning, statistical methods, financial knowledge and computer science to analyze financial data is common.

VIII. CONCLUSIONS

To meet the massive transaction demands of Yu'e Bao users, this paper proposed using both a linear model (YEB_ARIMA) and a nonlinear model (YEB_LSTM) to predict the future cash flow. Then, a YEB_Hybrid model integrating YEB_ARIMA and YEB_LSTM was constructed to enhance the prediction effect and calculate the fund reserve ratio. The model accuracies were quantitatively evaluated on a real dataset from Yu'e Bao from different perspectives, including a residual noise characteristics test, an LBQ test, and error calculation indexes. By making capital flow predictions, our method can effectively prevent liquidity risk and liquidity surplus to maximize business profits.

By utilizing this forecast model, fund managers could adjust their investment strategy in time to maximize benefit, which means that fund managers could reserve sufficient money when the money demand is large to avoid liquidity risks. A fund manager could also make more investments when the predicted money requirement decreases. Both of these circumstances could help fund managers avoid risks and make more income.

In the future, financial capital flow prediction should be applied by using the latest and newest dataset to train our model and verify its effectiveness. To better investigate how to dynamically monitor risk, the design and implementation of the YEB_LSTM network should also be optimized to improve its predictive capability [42]. Additionally, this analytical framework not only suits the Yu'e Bao product but also can be applied to other popular investment products in time series form. The capital flow framework could also be applied to mobile applications [43] for a better user experience [44].

REFERENCES

[1] Y. Yin, W. Xu, Y. Xu, H. Li, and L. Yu, "Collaborative QoS prediction for mobile service with data filtering and SlopeOne model," *Mobile Inf. Syst.*, vol. 2017, pp. 7356213:1–7356213:14, Jun. 2017.

[2] Y. Chen, S. Deng, H. Ma, and J. Yin, "Deploying data-intensive applications with multiple services components on edge," in *Mobile Networks and Applications*. Berlin, Germany: Springer, 2019.

[3] C. Harvie and T. Van Hoa, *The Causes and Impact of the Asian Financial Crisis*. Berlin, Germany: Springer, 2016.

[4] Q. Liu, F. Zhang, M. Mao, B. Xue, and Z. Lin, "An empirical study on factors affecting continuance intention of using Yu'e Bao," *Tehni ki Vjesnik*, vol. 25, no. 5, pp. 1414–1420, 2018.

[5] J. C. Gibbons, *Experiments in Quantitative Finance*. Abingdon, U.K.: Routledge, 2017.

[6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Kuala Lumpur, Malaysia: Pearson, 2016.

[7] S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On deep learning for trust-aware recommendations in social networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1164–1177, May 2017.

[8] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1054–1064, Apr. 2018.

[9] J. Gong, X. Ye, and Z. Liang, "An empirical study about the effect which 'Bao Bao' Internet monetary funds make on deposits in Chinese commercial banks," *Amer. J. Ind. Bus. Manage.*, vol. 6, no. 9, p. 993, 2016.

[10] P. Mathur, "Overview of machine learning in finance," in *Machine Learning Applications Using Python*. Berkeley, CA, USA: Apress, 2019, pp. 259–270.

[11] O. H. Hamid, N. L. Smith, and A. Barzanji, "'Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence," in *Proc. IEEE 15th Int. Conf. Ind. Inform. (INDIN)*, Jul. 2017, pp. 899–904.

[12] 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.

[13] W.-C. Wang, K.-W. Chau, D.-M. Xu, and X.-Y. Chen, "Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition," *Water Resour. Manage.*, vol. 29, no. 8, pp. 2655–2675, 2015.

[14] E. Paradis, "Moran's autocorrelation coefficient in comparative methods," R Found. Stat. Comput., Vienna, U.K., Tech. Rep., 2009, pp. 1–9. [Online]. Available: https://cran.r-project.org/web/packages/ape/vignettes/MoranI.pdf

[15] C.-W. Su, Z.-Z. Li, R. Tao, and D.-K. Si, "Testing for multiple bubbles in bitcoin markets: A generalized sup ADF test," *Jpn. World Economy*, vol. 46, pp. 56–63, Jun. 2018.

[16] M. Khashei and Z. Hajirahimi, "A comparative study of series arima/mlp hybrid models for stock price forecasting," *Commun. Statist.-Simul. Comput.*, pp. 1–16, May 2018. [Online]. Available: http://www.stat.ucla.edu/~frederic/415/F18/stockprice.pdf

[17] F. J. Pontes, G. F. Amorim, P. P. Balestrassi, A. P. Paiva, and J. R. Ferreira, "Design of experiments and focused grid search for neural network parameter optimization," *Neurocomputing*, vol. 186, pp. 22–34, Apr. 2016.

[18] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecasting*, vol. 32, no. 3, pp. 669–679, Jul./Sep. 2016.

[19] J. Kallsen and J. Muhle-Karbe, "The general structure of optimal investment and consumption with small transaction costs," *Math. Finance*, vol. 27, no. 3, pp. 659–703, 2017.

[20] S. Zhang, D. Zhou, M. Y. Yildirim, S. Alcorn, J. He, H. Davulcu, and H. Tong, "HiDDen: Hierarchical dense subgraph detection with application to financial fraud detection," in *Proc. SIAM Int. Conf. Data Mining Soc. Ind. Appl. Math.*, 2017, pp. 570–578.

[21] 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.

[22] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, 2018.

[23] D. T. Tran, A. Iosifidis, J. Kanniainen, and M. Gabbouj, "Temporal attention-augmented bilinear network for financial time-series data analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1407–1418, May 2019.

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[25] L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li, "Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3271–3278.

[26] Y. Yin, Y. Xu, W. Xu, M. Gao, L. Yu, and Y. Pei, "Collaborative service selection via ensemble learning in mixed mobile network environments," *Entropy*, vol. 19, no. 7, p. 358, 2017.

[27] M. H. Alobaidi, F. Chebana, and M. A. Meguid, "Robust ensemble learning framework for day-ahead forecasting of household based energy consumption," *Appl. Energy*, vol. 212, pp. 997–1012, Feb. 2018.

[28] H. Gao, S. Mao, W. Huang, and X. Yang, "Applying probabilistic model checking to financial production risk evaluation and control: A case study of Alibaba's Yu'e Bao," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 3, pp. 785–795, Sep. 2018.
[29] P. Bagchi, V. Characiejus, and H. Dette, "A simple test for white noise in functional time series," *J. Time Ser. Anal.*, vol. 39, no. 1, pp. 54–74, 2018.
[30] M. Brunnermeier, D. Palia, K. A. Sastry, and C. A. Sims, "Feedbacks: Financial markets and economic activity," Princeton Univ., Princeton, NJ, USA, Tech. Rep., 2017, pp. 1–64. [Online]. Available: https://scholar.princeton.edu/sites/default/files/markus/files/02bpss_draft.pdf
[31] I. Perera, J. Hidalgo, and M. J. Silvapulle, "A goodness-of-fit test for a class of autoregressive conditional duration models," *Econ. Rev.*, vol. 35, no. 6, pp. 1111–1141, 2016.
[32] T. Ye, S. Dianbo, and Z. Yichuan, "Statistical modeling analysis of fund flow of Yu'ebao," in *Proc. 4th Nat. College Students Stat. Modeling Contest*, 2015, pp. 1–32.
[33] O. Duru, R. Golan, and D. Quintana, "Computational intelligence in finance and economics [guest editorial]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, p. 13, Nov. 2018.
[34] V. Agarwal, S. Ruenzi, and F. Weigert, "Tail risk in hedge funds: A unique view from portfolio holdings," *J. Financial Econ.*, vol. 125, no. 3, pp. 610–636, Sep. 2017.
[35] L. Yongxin, "Discussing on securities investment fund hedging risks and countermeasures," in *Proc. Int. Conf. Inf. Manage., Innov. Manage. Ind. Eng.*, Sanya, China, Oct. 2012, pp. 94–97.
[36] Y. Sun, G. Aw, R. Loxton, and K. L. Teo, "Chance-constrained optimization for pension fund portfolios in the presence of default risk," *Eur. J. Oper. Res.*, vol. 256, no. 1, pp. 205–214, 2017.
[37] X. Zhang, H. Meng, J. Xiong, and Y. Shen, "Robust optimal investment and reinsurance of an insurer under Jump-diffusion models," *Math. Control Related Fields*, vol. 9, no. 1, pp. 59–76, 2019.
[38] H. Yan and H. Ouyang, "Financial time series prediction based on deep learning," *Wireless Pers. Commun.*, vol. 102, no. 2, pp. 1–8, 2018.
[39] Y. Yang, A. Kolesnikova, S. Lessmann, T. Ma, M.-C. Sung, and J. E. V. Johnson, "Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting," 2018, *arXiv:1812.06175*. [Online]. Available: https://arxiv.org/abs/1812.06175
[40] M. Liu, M. Huang, Y. Zhang, W. Feng, J. Lai, and X. Li, "Using deep residual networks to deal with financial risk control problems," in *Proc. ACAI*, New York, NY, USA, 2018, Art. no. 43.
[41] M. Ahmed, N. Choudhury, and S. Uddin, "Anomaly detection on big data in financial markets," in *Proc. IEEE/ACM ASONAM*, J. Diesner, E. Ferrari, and G. Xu, Eds. New York, NY, USA, 2017, pp. 998–1001.
[42] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Sep. 2018.
[43] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," in *Mobile Networks and Applications*. Berlin, Germany: Springer, 2019.
[44] S. Deng, Z. Xiang, J. Yin, J. Taheri, and A. Y. Zomaya, "Composition-driven IoT service provisioning in distributed edges," *IEEE Access*, vol. 6, pp. 54258–54269, 2018.

XIAOXIAN YANG received the Ph.D. degree in management science and engineering from Shanghai University, Shanghai, China, in 2017. She is currently an Assistant Professor with Shanghai Polytechnic University, China. Her research interests include business process management and formal methods.

SHUNYI MAO received the master's degree in computer application technology from the School of Computer Engineering and Science, Shanghai University, China, in 2018. Her research interests include deep learning, medical image processing, and model checking-based software verification. She is a CCF Member.

HONGHAO GAO received the Ph.D. degree in computer science from Shanghai University, in 2012. His research interests include service computing, model checking-based software verification, and sensors data application. He is an IET Fellow, BCS Fellow, EAI Fellow, CCF Senior Member, and CAAI Senior Member.

YUCONG DUAN received the Ph.D. degree in software engineering from the Institute of Software, Chinese Academy of Sciences, China, in 2006. He is currently a Professor and the Vice Director of Computer Science Department, Hainan University, China. His research interests include theoretical and empirical software engineering, model driven software development, and so on.

QIMING ZOU received the Ph.D. degree in machine manufacturing from Shanghai University, Shanghai, China, in 2015, where he is currently an Assistant Professor. His research interests include cloud computing and grid computing computer aided manufacturing.