Standing on the Shoulders of Long-Term: Short-Term Stock Prediction Based on Financial Report and Quotations

Hongchen Cao*
Shanghaitech University
Shanghai, China
caohch1@shanghaitech.edu.cn

Chenyi Xu*
Shanghaitech University
Shanghai, China
xuchy@shanghaitech.edu.cn

Chenbo Xi*
Shanghaitech University
Shanghai, China
xichb@shanghaitech.edu.cn

Ziyue Yang*
Shanghaitech University
Shanghai, China
yangzy@shanghaitech.edu.cn

ABSTRACT

With the help of the development of deep learning technology, stock price prediction is becoming one of the most popular research directions. A lot of efforts are invested into short-term or long-term data in the previous research, but no one has tried to combine short-term and long-term data to design an investment strategy. We propose a pipeline that utilizes both long-term and short-term data to output an investing strategy. We also conducted a simple simulated trading experiment, and the results show that the rate of return (ROR) of our method reached 12.51%.

CCS CONCEPTS

• **Applied computing** → *Economics*; • **Computing methodologies** → *Planning and scheduling*.

KEYWORDS

quantitative trading, stock prediction, deep learning

ACM Reference Format:
Hongchen Cao, Chenyi Xu*, Chenbo Xi*, and Ziyue Yang*. 2022. Standing on the Shoulders of Long-Term: Short-Term Stock Prediction Based on Financial Report and Quotations.

1 INTRODUCTION

Stock price prediction is a critical foundation in modern quantitative trading systems. The latest stock price prediction system relies on deep learning techniques, trained on stock transaction records in the past period, to suggest likely future prices given the current or historical quotations [3, 6, 12, 17].

We propose a pipeline that utilizes the combination of the long-term (i.e., financial report) and short-term (i.e., quotations) data to generate profitable investment strategies. Our approach uses long-term data to cluster stocks and analyzes their periodicity, and then exploit short-term data to predict future prices. By taking the advantage of two types of data, our approach can be better than traditional methods that rely only on fundamental quantitative analysis techniques or historical stock prices.

*All four authors contributed equally to this research.

2 BACKGROUND

2.1 Long-term & Short-term Stock

The long-term market is usually based on the year, and long-term investment focuses on the company's fundamental information and long-term development potential. This market is less affected by emergencies but requires an in-depth understanding of the investment target's field and the ability to analyze a large amount of relevant data on the investment target. Short-term trading takes days or even hours and minutes as the unit, it is more susceptible to unexpected events, but also more dependent on the analysis of historical transaction data. High-frequency trading is currently a very hot short-term trading method, which uses computer algorithms for price prediction and strategy formulation.

2.2 Financial Report

The purpose of preparing financial reports is to provide the users of financial statements with accounting information about the financial position, operating results, and changes in the financial position that are useful for making economic decisions.

3 RELATED WORK

3.1 Stock Prediction

Since the stock transaction record itself is a kind of time-series data, lots of works have studied how to apply existing models(e.g., MLPs [10], ARIMAS [4], LSTM [7]) to this field. Pang et al. [15] propose to use "stock vector" based on the development of word vector in deep learning, which achieves the accuracy of 52.5% for the Shanghai A-shares composite index. Kelotra et al. [9] integrate the rider optimization algorithm (ROA) and MBO to train the deep convolutional long short-term memory (Deep-ConvLSTM) model by their Rider-based monarch butterfly optimization (Rider-MBO) algorithm. Chung et al. [5] use the multi-channel convolutional neural networks (CNNs) to predict the fluctuation of the stock index and propose a method to systematically optimize the parameters for the CNN model by using a genetic algorithm (GA). However, since the stock market is very susceptible to events such as social trends, government policies, and natural disasters, only using historical stock transaction data for forecasting makes these methods often perform inferior to those combined data from multiple sources (e.g., news, equity relationship).

3.2 Financial Report Analysis

Typically, analysis of the Financial Report focuses on several aspects of the target company, among which the cash flow and competitiveness of the company are of the most interest to analysts. Also, analysts use Financial reports to predict risk and stock returns.

After BB [16], the accounting profession focuses on how to use the time series characteristics of earnings to predict future earnings. In the late 1980s, the accounting profession realized that using past earnings to predict future earnings was not accurate since used too little information. So many works turned their attention to using more information about financial statements or reports to predict earnings. Representatives among them are OU and PENMAN [1] and ACGP [2]. They decided to use information from reports as much as they can rather than presupposing differences in the ability of financial statement information to predict future earnings at first. Report items with significant predictability are selected through statistical analysis to forecast future earnings.

4 METHODOLOGY

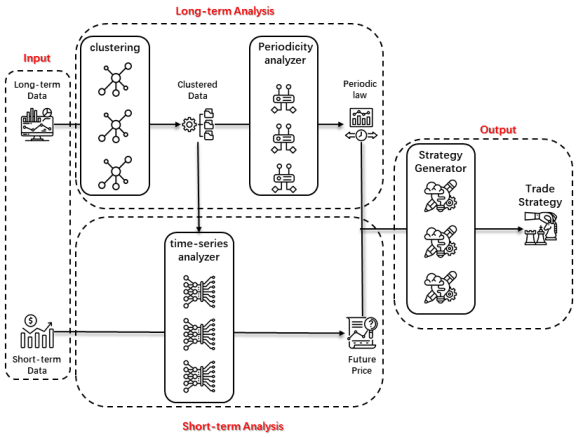


Figure 1: Overview of Our Pipeline

The financial report(Balance Sheet, Cash Flows Statement, etc.) is a financial data report that a stock company regularly announces to people. It not only has high credibility but also contains a lot of information such as the company's profitability, asset structure. The financial report is a kind of long-term data. For investors and the stock price prediction system, because of its large amount of information and reliable data sources, the financial report is of great reference value for evaluating the future price trend. The price trend of stocks in a short period is also an important factor affecting future stock prices[8, 11]. Based on the consolidated financial report data, the pipeline cluster the stocks and refers to short-term stock price trends to predict the future price of this part of the preferred stocks. These stock predictions assisted the investors in their decision of whether to buy a stock or not. Based on these sections, we design a pipeline to output an investing strategy.

As shown in Fig. 1, we design a pipeline to generate profitable investment strategies. We divide the process into three sections: **Long-term Analysis**, **Short-term Analysis** and **Trade Strategy Generation** (i.e., **Output**). The **Long-term Analysis** takes the long-term data(i.e., financial report) as an input and returns clustered stock data. The **Short-term Analysis** takes the historical quotations as its input and combines them with the result from the **Long-term Analysis** to give a prediction of the future price. Based on the results from both analyses, the pipeline finally outputs an investment strategy for the investors. We also analyze the periodicity of the stocks based on the clustered data to improve the final strategy.

4.1 Long-term Analysis

Lin et al. [13] found that different stocks have different trading patterns. After classifying the stocks into different patterns, short-term forecasts can be more accurate. Their method uses stock prices as evidence for classification. We think the difference in stock trading patterns comes from the difference in the company's structure, and the market's preference for a pattern of stock will affect the stock price. In this section, we use financial report data to cluster stocks to help future short-term forecasts.

Data Preprocessing.

We intercepted the companys' last four quarters' financial report data and took the average of the four values of each factor as input. We first take the logarithm of all the values because we think that the difference between one hundred million and one billion is close to one billion and ten billion. Another benefit of taking a logarithm is that it can reduce the impact of extreme values. After that, we standardize each factor dimension.

Let x be a two-dimensional financial report with shape $n * m$. Each row of x is a stock and each coloum is a factor. It will be calculated as $x_{ij} = ln(abs(x_{ij})) * sgn(x_{ij}), i \leq n, j \leq m$, where $sgn(x)$ indicates whether the symbol is positive or negative. Then $x = \frac{x-\mu(x)}{std(x)}$.

Factors selecting.

We will select the factors before clustering. The screening criterion is that the correlation of the selected factors is as low as possible. Because factors with low correlation do not mean that they have a great impact on stock prices (for example, randomly generated data has low correlation, but it is useless), we need to know exactly which factors have been selected. For such interpretability considerations, we do not intend to use algorithms such as PCA to reduce dimensionality, but to determine the Pearson correlation coefficient p between reference factors. The Pearson correlation is calculated by $r(x, y) = \frac{cov(x,y)}{\sqrt{std(x)*std(y)}}$.

We first select three factors f_1, f_2, f_3: total revenue, free cash flow, roe, and then select five of the remaining factors to minimize the sum of the correlation coefficients among the eight factors. That is to find $min(\sum_{i,j,i\neq j} abs(r(f_i, f_j)))$. We only select 8 factors, one is because 8 factors are sufficient for clustering, it is also due to calculation considerations. If more than 5 factors, we have to use intelligent algorithms such as SA or heuristic search, which may only get a locally optimal solution.

The final factors being chosen are as follow and the Pearson correlations between all factors are shown in Table. 1:

- Total revenue

- Free cashflow
- ROE: return on equity
- cap_rese_ps : capital reserve per share
- ocfps: net cash flow from operating activities per share
- ebit_ps: earnings before interest and taxes per share
- curr_to_debt : total current liabilities
- debt_to_eqt : debt to equity ratio

	Tot	Free	Roe	Cap	Ocfps	Ebit	Curr	Debt
Total revenue	1.0	0.103	0.305	-0.026	0.090	0.075	-0.055	0.021
Free cashflow	0.103	1.0	0.143	-0.119	0.137	0.022	0.068	-0.001
ROE	0.305	0.143	1.0	0.047	0.098	0.148	0.017	-0.231
cap_rese_ps	-0.026	-0.119	0.047	1.0	0.038	0.068	0.041	-0.048
ocfps	0.090	0.137	0.098	0.038	1.0	0.129	-0.017	0.009
ebit_ps	0.075	0.022	0.148	0.068	0.129	1.0	-0.011	-0.015
curr_to_debt	-0.055	0.068	0.017	0.041	-0.017	-0.011	1.0	-0.041
debt_to_eqt	0.021	-0.001	-0.231	-0.048	0.009	-0.015	-0.041	1.0

Table 1: Pearson correlation between factors

Clustering.
The Euclidean distance represents the absolute difference between two stocks, and the cosine similarity represents the similarity of the structure of the two stocks. It is possible to find similar stocks by using both distances. In our clustering method, we use cosine similarity + K-means as method 1 and then use the combination of cosine similarity and Euclidean distance + hierarchical clustering as method 2.

Method 1: Because cosine similarity and Euclidean distance have a mutual transformation relationship $euc = \sqrt{2 - cosine}$ for an unit vector, we can calculate the cluster centers of multiple points. So K-means is feasible for cosine similarity in section. K-means has the disadvantage that it is difficult to control the number of classes, which is not conducive to subsequent short-term predictions. We use divide and corner to overcome this shortcoming. As long as a category is larger than our demand, We divide it into two categories. Then we can control the max size of categories.

Method 2: We let L_1' is Euclidean distance and L_2' is cosine similarity. In order to make the smaller value is more similar, we let $L_2 = 1 - L_2' \in [0, 2]$. In order to ensure L_1 and L_2 are in the same range, we let L_1 normalized to $[0, 2]$ by $L_1 = \frac{2(L_1' - max(L_1'))}{max(L_1') - min(L_1')}$. Let $L(L_1, L_2) = L_1 * L_2 + k(L_1 + L_2)^2$ as the distace function, where k is a parameter and $k = 0.25$ in our cluster. This function has two characteristics. One is When the large distance increases, the distance between the two stocks will increase faster, which ensures that both two distances are not too large. Another is When the two distances are both small, the total distance is very small. This is because every term in this function is a quadratic term. After calculating the distance maxtrix, we use hierarchical clustering for clustering. This is beacuse it is hard to calculate cluster center by K-means.

Periodicity Analyzer.
Based on the observation that the price of some stocks tends to change periodically, we use the stocks price data to build a periodicity analyzer to judge whether a stock has periodicity or not.

We judge whether a stock has periodicity based on two criteria: whether the price of this stock shows a similar trend in two years, and whether the price has a similar trend for most of the period in the past. To evaluate the similarity of the price curve of the same stock in two different time intervals, we propose our stock periodicity evaluation algorithm based on Piecewise Linear Representation method [18] to calculate the derivative of two curves and evaluate the similarity of them. The details of the algorithm are shown in Alg. 1.

Those stocks which satisfy the conditions are considered to be periodic, and we record the periodic trend of the stock price.

ALGORITHM 1: PLR Periodicity Evaluation

```
Input : k is a custom threshold to judge whether two intervals are similar
DER1, DER2 = []
for a = 0 : dateNum − 1 do
    if Interval1[a + 1] − Interval1[a] >= 0 then
        DER1.append(1) # Get the derivative of the price in the first interval
    else
        DER1.append(−1)
    end if
end for
for b = 0 : dateNum − 1 do
    if Interval2[b + 1] − Interval2[b] >= 0 then
        DER2.append(1) # Get the derivative of the price in the second interval
    else
        DER2.append(−1)
    end if
end for
count = 0
for i = 0 : dateNum − 1 do
    if DER1[i] == DER2[i] then
        count = count + 1
    end if
end for
if count/(dateNum−1) >= k then
    return True # Periodic if the two derivatives are similar enough
end if
return False
```

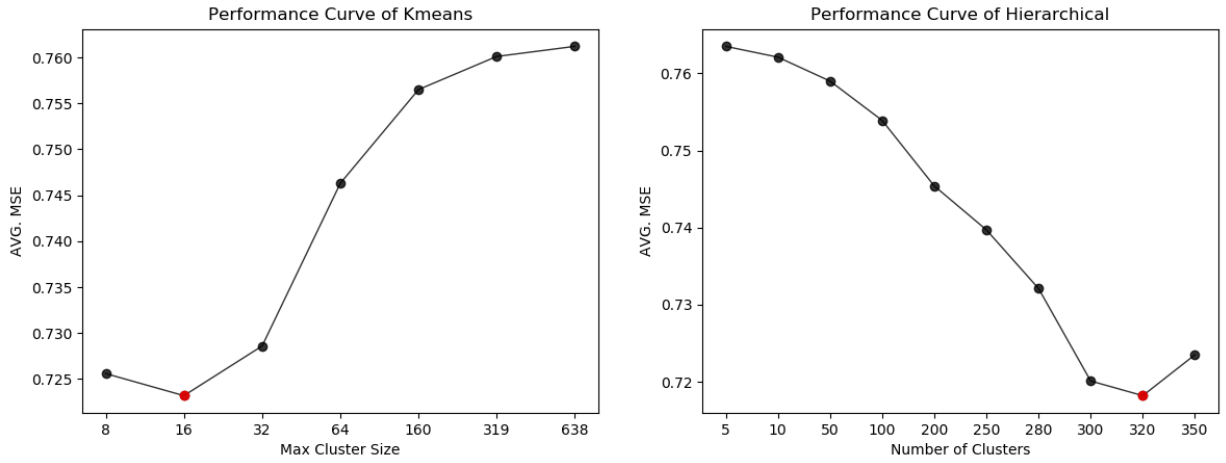
4.2 Short-term Analysis

We use a simple long short-term memory based deep learning model as the backbone of **Short-term Analysis**. Its input is the closing price of the past twenty trading days, and its output is the predicted closing price of the next trading day. Based on the clustering results obtained from the **Long-term Analysis**, we train an LSTM model for each cluster. **Long-term Analysis** ensures that the stocks in each cluster have similar characteristics and potential connections, which can help the model improve its predicting performance.

4.3 Trade Strategy Generation

The **Trade Strategy Generation** takes the results from both the **Long-term Analysis** and the **Short-term Analysis** as inputs. Through investment strategy generation, we can transform our prediction results into practical investment operation steps. In this process, the periodic characteristics of some stocks obtained through cluster analysis will further optimize the investment strategy and improve the results in the local scope.

This part can generate a strategy that can exploit out the predicted result of how much the market will change, rather than simply the direction it will shift. This allows us to base how much we invest on how certain we are of our prediction.



(a) K-means (b) Hierarchical

Figure 2: Performance curve of clustering algorithm

The Regression Strategy [14] is chosen because its good performance:

$$invest = \begin{cases} 100\% & \text{if change percent} > 1\% \\ p & -0.5\% \leq \text{change percent} \leq 1\% \\ 0\% & \text{if change percent} < -0.5\% \end{cases} \tag{1}$$

Here, *invest* is the percentage of our funds we use to buy stock, and the change percentage is computed by dividing the predicted change in the market tomorrow by the price today. The default value of p is 25% when not take periodic information into consideration.

5 EXPERIMENTS

5.1 Experiment Settings

Dataset: We use the dataset provided by SSE INFONET CO.,LTD. The dataset is mainly composed of two parts, historical stock price data, and financial report data. The historical stock price data contains 3192 stocks from January 2014 to November 2021. Some stocks have missing data for certain periods. To keep the dataset clean, we skip these dates instead of performing manual completion. The financial report contains a balance sheet, cash flow statement, income statement, and derivative financial indicators. Each indicator in the report contains data for each quarter from January 2014 to November 2021. We directly delete indicators that lack at least 30% of the data and use the median to fill in other missing data.

Evaluation Metrics: We use the mean squared error (MSE) and MSE improvement rate (MSEIR) to evluate the **Short-term Analysis** and rate of return (ROR) to evluate the output strategy. The detailed formulas of the three metrics are as follows:

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ MSEIR &= -\frac{MSE - \hat{MSE}}{\hat{MSE}} \\ ROR &= \frac{Cash_t - Cash_0}{Cash_0} \end{aligned} \tag{2}$$

5.2 Results of Long-term Analysis

As we mentioned in Sec. 4.1, we apply two different algorithms to cluster the stocks based on some specific features extracted from financial report data. As shown in Fig.2, we evaluate the relationship between the number of clusters and the prediction effect of two clustering algorithms.

For k-menas clustering (See Fig. 2(a)), we set the maximum size of clusters to 8(0.25% of the overall), 16(0.5%), 32(1%), 64(2%), 160(5%), 319(10%), 638(20%). The results show that when max cluster size equals to 16(0.5%), the best effect is achieved (i.e., $MSE = 0.723$).

For hierarchical clustering (See Fig. 2(b)), we set the number of clusters to 5, 10, 50, 100, 200, 250, 280, 300, 320, 350. The results show that when the number of clusters equals to 320, the best effect is achieved (i.e., $MSE = 0.718$).

By comparing the curves of the two clustering algorithms, we find that they have a similar trend. As the number of stocks in each cluster increases, the performance gradually deteriorates. The possible reason is that the larger the cluster size, the lower the average correlation between stocks in the cluster, which makes it difficult for the model to learn the correct future stock prices. This the problem may be solved by building a relationship graph of stocks and utilizing the graph neural networks to do clustering, but due to the limit of time and computing resources, we leave this part for future research.

5.3 Results of Short-term Analysis

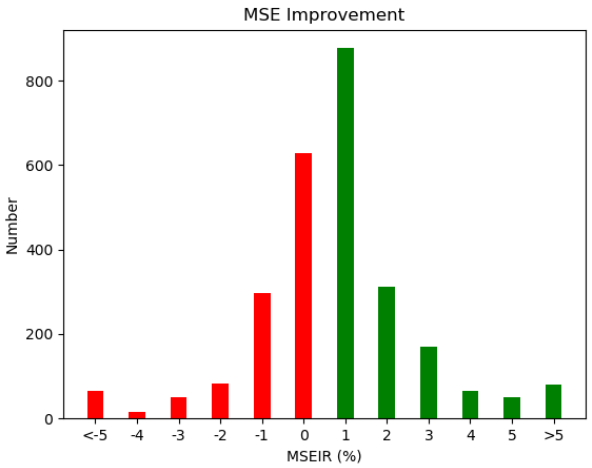


Figure 3: MSE Improvement

We calculate the MSEIR between the method combining **Long-term Analysis** (i.e., clustered data) and the method using only short-term data. Here we use the clustered result from **Long-term Analysis** based on hierarchical clustering and set the number of clusters to 300. As shown in fig.3, 57.684% of prediction results have improvements (i.e., MSE reduce). The average MSEIR is 0.720, the average MSEIR of the better part (i.e., green part in Fig. 3) is 1.378, the average MSEIR of the worse part (i.e., red part in Fig. 3) is −1.968.

We manually checked the stocks with MSE dropping, and we found that the average price of these stocks often differed significantly from the average price of other stocks in their clusters. A possible solution is to normalize the stock prices in each cluster, but it will compress the original price information, which will result in a decrease in model performance.

5.4 Results of Output Strategy

The evaluation of investment strategy is obtained through simulations. We start with enough money to buy one share of stock on the first day. Note that since we are only concerned with return results as percentages of starting money. At the start of each day, we invest according to some strategy based on prediction. At the end of the day, we sell all shares at the closing price.

For the omniscient strategy, we can know what exactly happens to the stock price tomorrow relative to today, and if it goes up, we buy the stock, and if it goes down, we do nothing. This is called the Theoretical maximum. At the same time, we chose the Basic LSTM method as the baseline to evaluate our approach.

We ran a simulation of each prediction method in terms of baseline basic LSTM, K-means clustering method, and Hierarchical clustering method. In order to evaluate our performance on period analysis, we use periodic information to further adjust our investment strategy parameters.

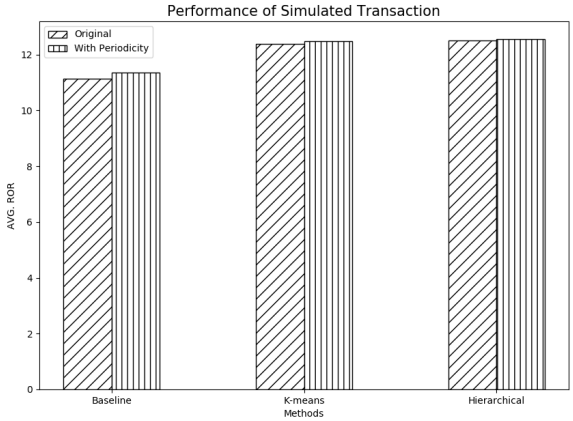


Figure 4: Performance of Simulated Transaction

We compare the ROR of the three group methods within one year, and the results are shown in Fig. 4. Regardless of the periodicity of stock prices, our clustering method enables us to achieve better returns than the basic LSTM method. Our two methods obtained average ROR of 12.39% and 12.51% respectively, while baseline is 11.14%.

We adjust our investment strategy by referring to the periodicity of stocks, especially the proportion of funds used to buy stocks in the current round, which describes how certain we are on the predicted results. As mentioned above, the setting of p-value reflects the adjustment of investment strategy through periodic information.

Taking periodic information into consideration, all methods above gain a higher ROR than before. The ROR of these methods is 11.35%, 12.48%, 12.56%, respectively. The baseline method, i.e. basic LSTM improved 0.21%, the highest of the three groups, at the same time, K-means clustering improved 0.09%, and Hierarchical clustering improved 0.05%. This partly illustrates the validity of periodicity. However, as is the case, many stocks are weak or non-cyclical, which also explains why periodic information can't get significantly better results after doing clustering. One explanation is that similar stocks tend to be cyclical at the same time, and they tend to be clustered into the same group.

6 CONCLUSION

We design and implement a stock price forecasting system that utilizes both long-term and short-term data and further generates investment strategies. Experiments on the dataset containing 3192 stock information show that our pipeline can achieve a ROR of 12.51%. We have also conducted analysis and experiments including clustering algorithms and periodicity search methods. Limited by time and computing resources, we cannot further explore how to embed more complex deep learning techniques into our pipeline to achieve better performance, but our results prove that the combination of long-term data and short-term data to predict stock prices and generate investment strategies is feasible and effective.

REFERENCES

[1] Ou Jane A. and Stephen H. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics* 11, 4 (1989), 295–329.

[2] Ou Jane A. and Stephen H. Penman. 1999. Financial analysis, future earnings and cash flows, and the prediction of stock returns: evidence for the UK. *Accounting and Business Research* 29 (1999), 281–298.

[3] Wei Chen, Haoyu Zhang, Mukesh Kumar Mehlawat, and Lifen Jia. 2021. Mean-variance portfolio optimization using machine learning-based stock price prediction. *Appl. Soft Comput.* 100 (2021), 106943.

[4] Jiajia Cheng, Huiyun Deng, Guang Sun, Peng Guo, and Jianjun Zhang. 2020. Application of ARIMA Model in Financial Time Series in Stocks. In *Proceedings of ICAIS*, Vol. 12239. 232–243.

[5] Hyejung Chung and Kyung-shik Shin. 2020. Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction. *Neural Comput. Appl.* 32, 12 (2020), 7897–7914.

[6] Arvand Fazeli and Sheridan K. Houghten. 2019. Deep Learning for the Prediction of Stock Market Trends. In *Proceedings of Big Data*. 5513–5521.

[7] Vasilis Karlis, Katerina Lepenioti, Alexandros Bousdekis, and Gregoris Mentzas. 2021. Stock Trend Prediction by Fusing Prices and Indices with LSTM Neural Networks. In *Proceedings of IISA*. 1–7.

[8] Ahmad Kazem, Ebrahim Sharifi, Farookh Khadeer Hussain, Morteza Saberi, and Omar Khadeer Hussain. 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Appl. Soft Comput.* 13 (2013), 947–958.

[9] Amit Kelotra and Prateek Pandey. 2020. Stock Market Prediction Using Optimized Deep-ConvLSTM Model. *Big Data* 8, 1 (2020), 5–24.

[10] Mehdi Khashei and Zahra Hajirahimi. 2019. A comparative study of series arima/mlp hybrid models for stock price forecasting. *Commun. Stat. Simul. Comput.* 48, 9 (2019), 2625–2640.

[11] Salim Lahmiri. 2018. Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression. *Appl. Math. Comput.* 320 (2018), 444–451.

[12] Bo Li and Li-feng Li. 2021. Stock Prediction Based on Adaptive Gradient Descent Deep Learning. In *Proceedings of AINA*, Vol. 225. 51–62.

[13] Hengxu Lin, Dong Zhou, Weiqing Liu, and Jiang Bian. 2021. Learning Multiple Stock Trading Patterns with Temporal Routing Adaptor and Optimal Transport. In *KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1017–1026.

[14] Te Braak P. Nayak R. 2007. Temporal pattern matching for the prediction of stock prices. *Australian Computer Society and AIDM 2007* (2007), 95–104.

[15] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, and Victor Chang. 2020. An innovative neural network approach for stock market prediction. *J. Supercomput.* 76, 3 (2020), 2098–2118.

[16] Ball R and Brown P. 1968. An empirical evaluation of accounting income numbers. *Journal of accounting research* (1968), 159–178.

[17] Heyuan Wang, Shun Li, Tengjiao Wang, and Jiayi Zheng. 2021. Hierarchical Adaptive Temporal-Relational Modeling for Stock Trend Prediction. In *Proceedings of IJCAI*. 3691–3698.

[18] Xiyang Yang, Jing Zhang, Fusheng Yu, and Zhiwei Li. 2019. Mathematical Programming for Piecewise Linear Representation of Discrete Time Series *(Advances in Intelligent Systems and Computing, Vol. 1075)*. 157–167.