Testing the Market Efficiency by LSTM and SVM

Tengyue Zhang[1,*]

[1] *Sunwah International Business School, Liaoning University, Liaoning, China*
[*] *Corresponding author. Email: 2016123862@jou.edu.cn*

ABSTRACT

As an essential part of risk investment and a microcosm of the national economy, predicting the stock market's change accurately and efficiently becomes extremely important. The purpose of this paper is to evaluate the accuracy of SVM and LSTM models to judge whether the Efficient Markets Hypothesis (EMH) is correct or not by predicting the typical stock indexes of the relatively mature American stock market and the gradually mature Chinese stock market. Therefore, this article applies the Kaggle Data Set to predict the stock price of S&P 500 and SSEC from January 01, 2013 to January 01, 2018 by using both the LSTM model and the SVM model. First, this paper compares the predicted trends with the actual trend respectively. Second, this paper compares the two stocks and concludes the efficiency of markets in different countries. Third, this paper analyzes the influence of different policies on stock market fluctuation to explain the unpredictable change in the stock market. Finally, according to the results, the statistically significant conclusions are drawn that LSTM is more stable and accurate than SVM in the stock indexes prediction and American stock market is more effective than the Chinese stock market. Therefore, relevant forecasters can be more inclined to use LSTM model when making predictions.

Keywords: *stock index prediction, SVM, LSTM, market efficiency*

1. INTRODUCTION

Since the establishment of the stock market, there has been substantial active trading activities. The common consensus of the stock market is that it changes from moment to moment. As the financial market is constantly changing, the financial derivatives are also constantly changing, which leads to the uncertainty of investment returns. Besides, the stock market variations, for instance, short selling, may result in various perils, such as stock market crashes. What's more, stock analysis costs substantial time and money to form investors' own investment experience, which is highly subjective. [1]

However, applying computers to spontaneously study and analyze historical data in the market and form models to predict the future trend of stocks can not only reduce the cost of manual learning but also avoid investors being affected by subjective market sentiment. Therefore, many scholars are also constantly exploring the role of machine learning in the stock market which is why many quantitative hedge funds and stock trading firms hire millions of Mathematics PHDs and Physics PHDs to find the law of stock migration using various models. [2] Of hundreds of models which we are familiar with, the Long Short-term Memory (LSTM) model and the Support Vector Machine (SVM) are commonly used. The LSTM model uses known stock market data to analyze the ups and downs of a given period and predict whether a stock will rise or fall next. The problems of gradient disappearance and gradient explosion in long sequence training could be solved by its advantages in sequence modeling and long-term memory. [3] The SVM model is based on nonlinear mapping theory. The core of the support vector machine method is the idea of the optimal hyper plane of feature space partition and maximization of classification margin. It is crucial in classification decision making.

To make full use of our knowledge, this paper makes the following arrangements. First, the paper briefly introduces the Efficient Markets Hypothesis (EMH) to draw forth the idea that investors are difficult to obtain excess profits higher than the market average especially by analyzing past prices. Second, the paper collects the data of stock price from the Kaggle Data Set and leverages LSTM and SVM models to predict the stock price changes of S&P 500 and SSEC during the period from January 01, 2013, to January 01, 2018. Then, the paper finds out the LSTM model is much better than the SVM model by comparing different prediction results with the actual stock price. Finally, the paper analyses the

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 584–590, 2022.
https://doi.org/10.2991/978-94-6463-036-7_86

market efficiency by comparing the stock trend charts of S&P 500 and SSEC. The paper also analyses the different macroeconomic policies of the two countries to make a conclusion that the market of the USA is more efficient than the market of China and it is quite important to make a fundamental analysis.

This paper is constructed as follows. Section 2 reviews the literature. Section 3 shows the data and the methods. Section 4 compares the results. Section 5 draws the conclusion.

2. LITERATURE REVIEW

The Efficient Markets Hypothesis (EMH) was put forward by Fama in 1965. It states that a security market is efficient if its prices fully reflect all available information. This assumption assumes that people are rational economic agents. The weak, semi-strong, and strong form hypotheses are the three forms of efficient market hypothesis, and the three forms are progressive. [4]

The weak form hypothesis holds that market prices have fully reflected all historical security price information, including various prices and quantities of stocks. The semi-strong form hypothesis indicates that a company's condition can be fully represented by prices, including stock prices and trading volumes mentioned in the first hypothesis, as well as macroeconomic aspects and financial reports that can reflect a company's operating and financial conditions. The last strongly efficient market hypothesis reveals that both publicly available information and undisclosed information are already reflected in prices. [5]

However, since there is no specific measurement standard for market effectiveness, different scholars have tested market effectiveness empirically through different methods. Autocorrelation test, run test, unit root test, and variance ratio were used to test the weak-form market efficiency of stock market returns in Asia-pacific markets. [6] They concluded that monthly prices did not follow a random walk across all countries in the Asia-Pacific region, meaning investors could profit from the arbitrage process.

Many existing works on the stock predictions are based on Linear Regression. [7] [8] It can tell on average, what has been the past trend. The prediction will be quite accurate only if the future trend happens to be the same as the past trend. In reality, it is always possible that the trend reverses which the linear forecast can never reveal. Therefore, more complex models should be chosen in order to cope with the variability in the actual data. [9]

3. DATA AND METHOD

3.1. Data

The statistics of this article are based on the Kaggle Data Set (http://www.kaggle.com) to predict the stock price of S&P 500 and SSEC from January 01, 2013 to January 01, 2018 and build a moving window that uses the first 5 days' data to predict the sixth day's data. Besides, 80% of the overall data is taken as the training set of the model and 20% is taken as the test set of the model. To predict the accuracy of LSTM and SVM, the paper uses the Mean Square Error loss function and the Adam optimizer. At the same time, Root Mean Square Error (RMSE) is used to evaluate the accuracy of prediction.

Through visualizing the data in Table **1**, data characteristics including mean, standard deviation, minimum and maximum. When it comes to these characteristics, the index value of SSEC is consistently higher than that of S&P500.

Table 1. Data description

	S&P500	SSEC
Mean	1698.5818	2926.0284
Standard deviation	546.7881	758.1871
Minimum	676.5300	1706.7000
Maximum	2930.7500	6092.0600

3.2. Method

Since the traditional simple linear regression model is not effective in stock prediction, this article selects SVM and LSTM models, which have outstanding performance in the existing research results. [10]

To propose the Support Vector Machine Algorithm, the following equation (1) represents the mathematics involved behind SVM, also known as the primal formula which subjects to formula (2) (3) (4) (5): [11]

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \tag{1}$$

Subjects to:

$$\forall n: y_n - (x_n'\beta + b) \le \varepsilon + \xi_n \tag{2}$$

$$\forall n: (x_n'\beta + b) - y_n \le \varepsilon + \xi_n^* \tag{3}$$

$$\forall n: \xi_n^* \ge 0 \tag{4}$$

$$\forall n: \xi_n \ge 0 \tag{5}$$

ξ_n and ξ_n^* are slack variables. The constant C is used to control the penalty which lies outside the epsilon margin (ε), known as the box constraint, a positive numeric value.

LSTM model is a special type of Recurrent neural network (RNN). It had a long history and was first proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. It is a special algorithm to train the data with time-series information, that is, these data are not only arranged in ascending order of time, but also have a strong connection with the data before and after. The LSTM model can be summarized as the following equations (6) (7) (8) (9) (10) (11). For more details, please refer to Hochreiter. [12]

$$f_t = \sigma_g(W_f \times x_t + U_f \times h_{t-1} + b_f) \quad (6)$$

$$i_t = \sigma_g(W_i \times x_t + U_i \times h_{t-1} + b_i) \quad (7)$$

$$o_t = \sigma_g(W_o \times x_t + U_o \times h_{t-1} + b_o) \quad (8)$$

$$c'_t = \sigma_c(W_c \times x_t + U_c \times h_{t-1} + b_c) \quad (9)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t \quad (10)$$

$$h_t = o_t \cdot \sigma_c(c_t) \quad (11)$$

f_t is the forget gate. i_t and o_t represents the input gate and output gate respectively. c_t is the cell state. h_t is the hidden state.

LSTM model uses known stock market data to analyze the ups and downs of a given period of time and predict whether the stock will rise or fall next. LSTM model is good at sequence modeling and long-term memory and is easy to implement. [13]

With the aim of the optimal hyperplane of feature space partition, the SVM model is based on the theory of nonlinear mapping, the optimal hyperplane of feature space partition, and the idea of maximizing classification margin. The support vector is the key to the SVM classification decision. SVM model has strict theoretical foundations which basically do not include probability measures and the law of large numbers; therefore, it is different from the existing statistical methods. Moreover, it uses deduction as a substitute for induction in the traditional process. SVM model also achieves efficient "transduction inference" from training samples to prediction samples, greatly simplifying the usual classification and regression problems.

After the data is preprocessed, this paper normalizes the data first. Second, the article uses the historical data of the first 5 days to predict the data of the sixth day and build a moving window 80% of the overall data is taken as the training set of the model and 20% is taken as the test set of the model. Third, this paper builds the model using the Mean Square Error as the loss function and the Adam as the optimizer. Finally, predict data and reverse normalization to restore the data. Root mean squared error (RMSE) is used as the evaluation of the accuracy of prediction. **Figure 1** shows the process mentioned above.

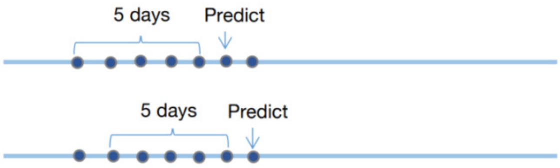

Figure 1. Moving window

4. RESULTS

To find out which model can predict the alteration of stock price more accurately we based on the Kaggle Data Set to predict two different stocks: S&P 500 from the USA and SSEC from China from January 01, 2013 to January 01, 2018 using both the LSTM model and the SVM model. After the training of the sample set, the predicted performance of the two machine learning models in the test set is shown in the table. As shown in the following **Table 2** and **Figure 2 3 4 5**, the blue line shows the real stock price, and the orange line shows the prediction of the stock index by SVM or LSTM mode. [14]

From the line graphs above we can clearly see that the MSE of LSTM in S&P 500 is 3.99e-0.5 and MAE is 0.0046 while the MSE of SVP in S&P 500 is 0.000149 and MAE is 0.0105. For SSEC the MSE of LSTM is 0.00027 and MAE is 0.01303 while the MSE of SVM is 0.01689 and the MAE is 0.12716. [15] Since the LSTM's numerical value of MSE and MAE are far lower than SVP'S that indicates the prediction ability of SVM fluctuates greatly, while the prediction ability of LSTM fluctuates less. That's because quadratic programming is used to solve support vectors. However, we will have to calculate m-order matrices in quadratic programming. When it comes to a large number of samples, machine memory and computing time consumption are inevitable for matrix storage and calculation. Therefore, the SVM algorithm is incompatible with large-scale training samples. What's more, the traditional support vector machine algorithm only applies to binary classification. However, in practice, the multi-class classification problems need to be solved in the data mining application. Since the stock market has thousands of millions of data, solving multiple classification problems becomes extremely difficult when using SVM. But for LSTM, it is suitable for learning time series and has a certain memory effect on time series sensitive problems and tasks. The LSTM model also adds special units of memory cells, such as accumulators and gated neurons. Therefore, in the next time step, LSTM will have a weight connected to itself in parallel and at the same time copy the real value of its own state and the accumulated external signal. This special implicit unit increases the network storage and can learn and store information for a long time. That is why the prediction effect of LSTM is somehow awesome. [16]

Since the fluctuation of stock price can reflect the current economic situation of the whole country directly, it is accepted by most people that the fluctuation of stock price can reflect all available information. [17] Through LSTM's analytic data shown above and the structure and subsequent variance ratio test, we can clearly see that S&P 500 and the US market represented by it are far better than SSEC and the Chinese market represented by it, in terms of overall size and market size. [18] In order to find out the reason for this, the paper shows the line graph in **Figure 6.**

The black line shows the money the Fed has pumped into the economy, which also represents the degree of quantitative easing. The green line represents the overall U.S. stock market. [19] So we can see that from 2008 to 2020, the U.S. stock market continued to rise rapidly with the central bank's injection of capital. [20] Since the stock market rises and falls in a positive correlation with the economy, just like where there is a lot of money, especially where there are a lot of rich people, prices will continue to rise, the continues raising of the US stock price is inseparable from the central bank's capital injection. [21] This paper analyzes three reasons why the US stock market had a 12-year bull market from 2008 to 2020. The first reason is quantitative easing and zero interest rates. The second reason is tax relief, and the last reason is that capital continues to flow to the U.S. due to weak markets elsewhere after the financial crisis. Because of the large amount of money in the U.S. economy and the general confidence in the U.S. stock market, people would buy stocks once the U.S. stock market fell, and thus the overall bear market in the U.S. stock market did not occur in the past 12 years. [22]

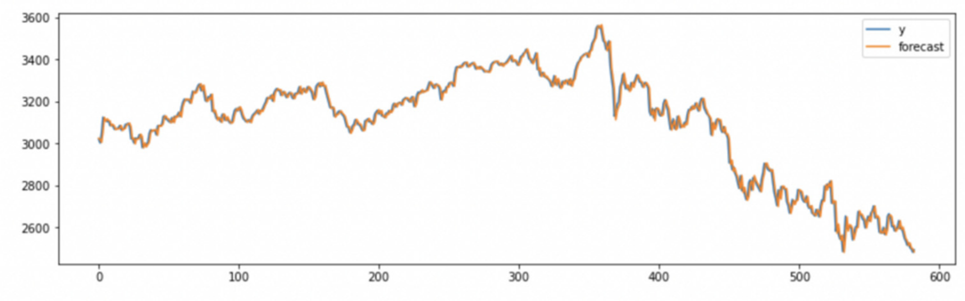

Figure 2. Prediction of S&P500 using LSTM

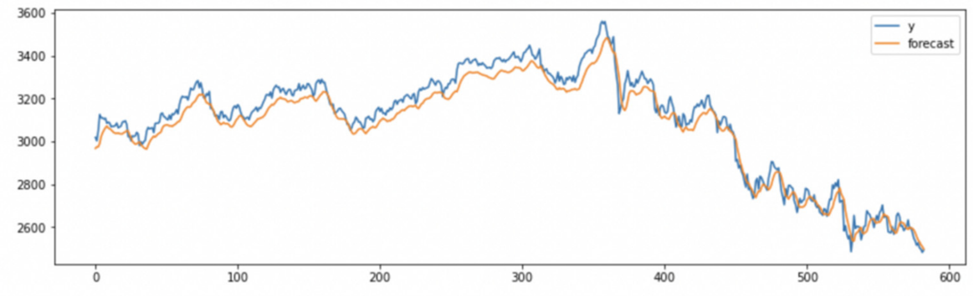

Figure 3. Prediction of S&P500 using SVM

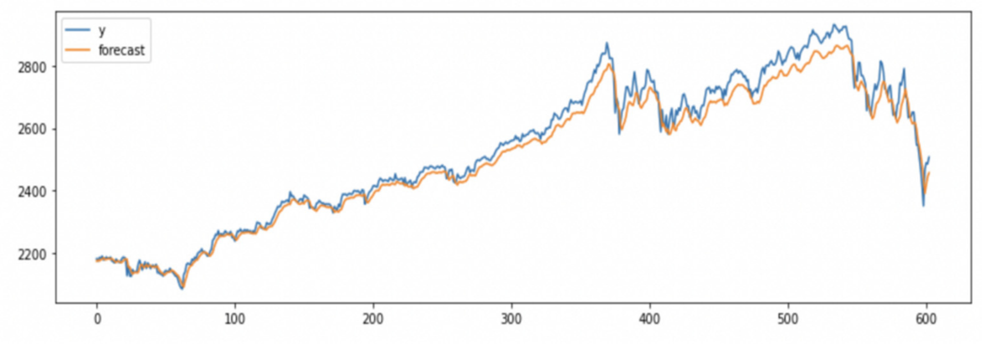

Figure 4. Prediction of SSEC using LSTM

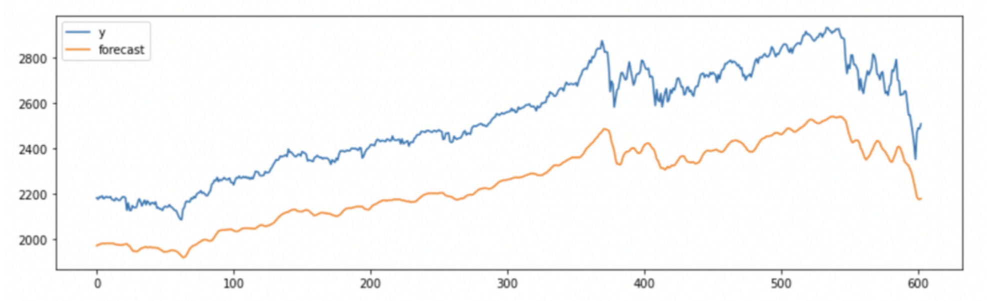

Figure 5. Prediction of SSEC using SVM

Table 2. RMSE of results

Stock index type	RMSE of LSTM	RMSE of SVM
S&P 500	0.016	0.130
SSEC	0.006	0.012

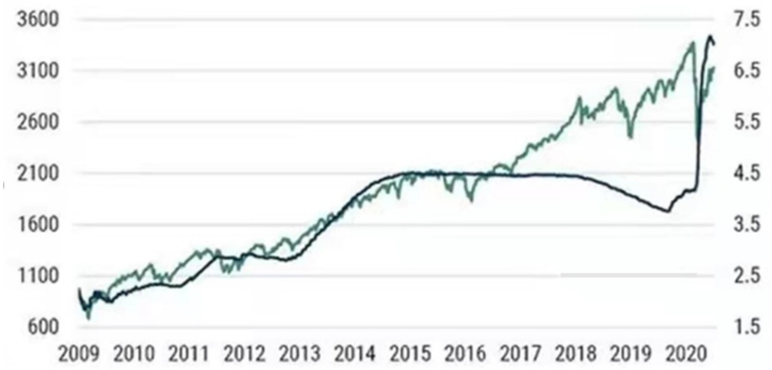

Figure 6. The US Central bank injection and the stock price chart

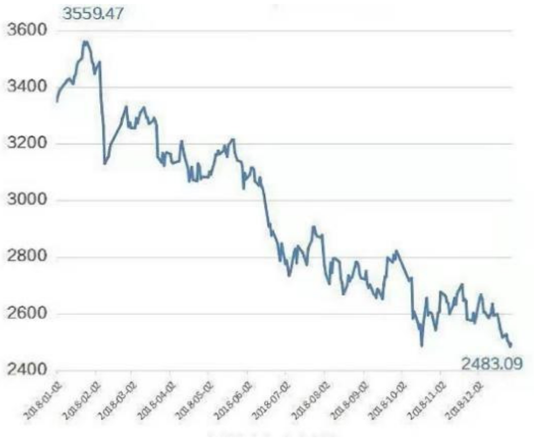

Figure 7. Chinese stock price chart in 2018

On the contrary, as **Figure 7** shows, since China began the deleveraging policy in 2017, there is less money in the overall domestic economy. So, 2018 has been a bear market in China. [23]

5.CONCLUSION

In this paper, first, the stock price of S&P500 and SSEC from 2013 to 2018 are analyzed by using both the LSTM model and the SVM model. Through the analysis of the above two models, both the SVM and LSTM models can predict the stock market movements, but with regard to accuracy and efficiency of prediction, the LSTM model is far better than the SVM model due to its special units of memory cells and powerful network storage. Second, the paper compares the stock price of S&P500 and SSEC, at the same time discovering that S&P 500 representing the US market is more efficient than SSEC representing the Chinese market, not only in its cash flow but also in its market size. Third, through the comparison with the random walk model, the paper finds out that neither the prediction of LSTM nor the prediction of SVM performs well, and shows that there is much information in historical market data as input variables that can improve future price expectations. The paper explains the phenomenon by showing the stock price is random walking and there are many reasons to cause the fluctuation of the stock price it is very important to focus on not only technical analysis but also fundamental analysis. For example, the national policy such as quantitative easing, tax relief, and also deleveraging, what’s more, the impact of public opinion on the industry, like when Max announced that the world could buy Tesla with bitcoin, causing bitcoin stocks to jump.

However, there are still several aspects of this paper that need to be improved. First, the time interval of the selected training set is not long enough, so there are few valid out-of-sample test results. In addition, because the application of improved Support Vector Machine in inventory forecasting has not been explored, the ideal forecast effect has not been achieved. The results of support vector machines may be better if they are combined with some proven algorithms. In addition, the input characteristic variables used in the machine learning model in this paper are only market indicators, which may be difficult to extract information for the machine learning model. If more factors such as corporate financial indicators, macro and industry indicators can be provided as input variables, more conclusions can be drawn compared with the empirical test results in this paper.

REFERENCES

[1] Zhang Haiying, Liang Qiaomei, Li Siheng... & Wu Qingqiang.(2020).Research on Stock Prediction Model Based on Deep Learning. *Journal of Physics: Conference Series* (2). doi:10.1088/1742-6596/1549/2/022124.

[2] Tuarob Suppawong, Wettayakorn Poom, Phetchai Ponpat, Traivijitkhun Siripong, Lim Sunghoon, Noraset Thanapon & Thaipisutikul Tipajin.(2021).DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction. *Financial Innovation* (1). doi:10.1186/S40854-021-00269-7.

[3] Mehtab, S., Sen, J., & Dutta, A. (2020, October). Stock price prediction using machine learning and LSTM-based deep learning models. In *Symposium on Machine Learning and Metaheuristics Algorithms, and Applications* (pp. 88-106). Springer, Singapore.

[4] Schwartz Robert A..(1970).Efficient Capital Markets: A Review of Theory and Empirical Work: Discussion. *The Journal of Finance*(2). doi:10.2307/2325488.

[5] Gili Yen & Cheng-few Lee.(2008).Efficient Market Hypothesis (EMH): Past, Present and Future. *Review of Pacific Basin Financial Markets and Policies*(2). doi:10.1142/S0219091508001362.

[6] Hamid, Kashif and Suleman, Muhammad Tahir and Ali Shah, Syed Zulfiqar and Imdad Akash, Rana Shahid, Testing the Weak Form of Efficient Market Hypothesis: Empirical Evidence from Asia-Pacific Markets (February 7, 2017). Available at SSRN: https://ssrn.com/abstract=2912908 or http://dx.doi.org/10.2139/ssrn.2912908

[7] Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017, April). Stock market predication using a linear regression. In *2017 international conference of electronics, communication and aerospace technology (ICECA)* (Vol. 2, pp. 510-513). IEEE.

[8] Seethalakshmi, R. (2018). Analysis of stock market predictor variables using linear regression. *International Journal of Pure and Applied Mathematics*, *119*(15), 369-378.

[9] Altay, E., & Satman, M. H. (2005). Stock market forecasting: artificial neural network and linear regression comparison in an emerging market. *Journal of Financial Management & Analysis*, *18*(2), 18.

[10] Song Donghwan, Baek Adrian Matias Chung & Kim Namhun.(2021).Forecasting Stock Market Indices Using Padding-Based Fourier Transform Denoising and Time Series Deep Learning Models. *IEEE ACCESS*. doi:10.1109/ACCESS.2021.3086537.

[11] Fan, R.E. , P.H. Chen, and C.J. Lin. "A Study on SMO-Type Decomposition Methods for Support Vector Machines." *IEEE Transactions on Neural Networks,* Vol. 17:893–908, 2006.

[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[13] De-hua LIU & Jing-jing WANG. (2018).A PCA-LSTM Model for Stock Index Prediction..(eds.)*DEStech Transactions on Engineering and Technology Research*(pp.)..

[14] Cumhur Buguk & B Wade Brorsen.(2003).Testing weak-form market efficiency: Evidence from the Istanbul Stock Exchange. *International Review of Financial Analysis*(5). doi:10.1016/S1057-5219(03)00065-6.

[15] McMillan David G..(2021).Forecasting U.S. stock returns. *The European Journal of Finance*(1-2). doi:10.1080/1351847X.2020.1719175.

[16] Bruno Miranda Henrique,Vinicius Amorim Sobreiro & Herbert Kimura.(2018).Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data*

Science(3). doi:10.1016/j.jfds.2018.04.003.

[17] Alves Luiz G. A., Sigaki Higor Y. D., Perc Matjaž & Ribeiro Haroldo V..(2020).Collective dynamics of stock market efficiency. *Scientific Reports* (1). doi:10.1038/s41598-020-78707-2.

[18] Sudhakara Reddy Syamala & Kavita Wadhwa.(2020).Trading performance and market efficiency: Evidence from algorithmic trading. *Research in International Business and Finance*. doi:10.1016/j.ribaf.2020.101283.

[19] MENAFN.(2020).Experienced Equity Research Analyst Launches Online Platform to Analyze the US Stock Market. *M2 Presswire*.

[20] Abbas Khan,Muhammad Yar Khan,Abdul Qayyum Khan,Majid Jamal Khan & Zia Ur Rahman.(2021).Testing the weak form of efficient market hypothesis for socially responsible and Shariah indexes in the USA. *Journal of Islamic Accounting and Business Research*(5). doi:10.1108/JIABR-02-2020-0055.

[21] Bentes Sónia R..(2021).On the hysteresis of financial crises in the US: Evidence from S&P 500. *Physica A: Statistical Mechanics and its Applications*. doi:10.1016/j.physa.2020.125583.

[22] Abdelkader Boudriga & Dorsaf Azouz Ghachem.(2018).Does American Stock Market React Differently to Expected Versus Surprise Ratings During Crisis Period? The Case of the 2008 Worldwide Financial Crisis. *International Journal of Accounting and Financial Reporting*(3). doi:10.5296/ijafr.v8i3.13587.

[23] Ting Zhang & Yiqi Zhuang. (2020).Research on the impact of Fintech event on Chinese commercial banks' stock price. *International Journal of Wireless and Mobile Computing* (3).