Comparing the Efficiency of Stock and Index Price Prediction Between FNN and LSTM Scenarios

Yigong Yuan*
* Corresponding author: 191098312@smail.nju.edu.cn

Department of Finance and Insurance, Nanjing University, Beijing, China

Abstract—Predicting the trend of changes in stock prices serves as a crucial role in the quantitative investment industry, and few previous empirical analysis studies choose the underlying assets from Hong Kong Stock Exchange. This study compares the performance of two machine learning approaches (i.e., FNN and LSTM) in the interpretation accuracy of two Hong Kong Internet technology industry companies and two related market indices, for the sake of exploring a better way to explain the Hong Kong stock price movement. Ten technical analysis factors were selected as inputs to the model. According to the analysis, after selecting the best model parameters and training times, both FNN and LSTM models have high accuracy in predicting Hong Kong stock data (82.76% and 90.85% respectively). In comparison, LSTM is more suitable for analyzing Hong Kong Internet technology Model selection for stock movement trends. These results shed light on guiding further exploration of stock price prediction in terms of the state-of-art neural network techniques.

Keywords-FNN; LSTM; HongKong Stock; Internet Industry.

1 INTRODUCTION

Stock price prediction has always been a difficult issue due to the stochastic fluctuation of markets. To analyze potential future values, two types of method are generally accepted by investors. One is fundamental analysis by which investors measure the intrinsic value of the stock by examining the related macro economical, industrial and corporate financial factors. The other method is technical analysis, which relies on statistical trends and indicators to predict the future movement of stock price. Technical analysis is based on the efficient market hypothesis and the assumption that history will repeat in the future. Invented by Malkiel and Fama in 1970, the efficient market hypothesis means that the prices of underlying assets reflect all information of the market and therefore continuously beating the market is impossible. It has become the logical base of the prediction of technical analysis. Since researchers can hardly gather and evaluate all information in the markets ranging from political affairs to corporate accounting changes due to the information asymmetry, it is more reasonable using the stock price and volume that has already been efficiently measured, reflecting all information in the market. Beyond this, the hypothesis of history repeating itself makes the prediction possible.

Although linear regression model has always been a generally accepted way of technical analysis, the non-linear machine learning techniques is swiftly gaining popularity in the investment community and is assumed to have a better performance owing to the applicability to non-

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328401

stationary time series stock market data. Artificial neural network (ANN) is one of the models that has been most widely used to predict financial data.

This study focused on the comparison of three types of ANN model, feedforward neural network (FNN) and long short-term memory (LSTM), a special form of recurrent neural network (RNN), with 10 parameters introduced to analyze the close price for two stock price indexes that indicate the general market and internet industry performances (HSI and HSIII) and two individual internet technology stocks (Tencent and SMIC). The prediction performance on test dataset will be measured through ADAM and RMSE approach. Both stocks and indices are highly voluminous and vehemently traded in, so they reflect China's internet industry as a whole.

In the practice of financial time series data prediction, ANN model is currently the most effective means. Artificial neural network is a series of algorithms that recognizing the underlying relations of data through a process that mimic human brain operation. It has firstly been used in natural science study then introduced into financial analysis. Kim and Han proposed using ANN model to predict stock price index combining with the GAs to improve the accuracy and reduce feature dimensions, proved to have a better prediction performance [1]. Precious studies have shown that ANN model has a generously higher accuracy in predicting individual stock and stock price index. Patel et al. stated that ANN outperformed other machine learning approaches (including SVM, random forest and Naive-Bayes) on both continuous and trend input approaches for predicting daily price of Indian market [2]. Vijha et al. have also compared the performance of ANN and random forest in forecasting the close price of 4 American stocks, suggesting that ANN is more efficient under the standard of RMSE and MAPE [3]. The comparison among different types of ANN model has also been carried out. Guresen et al. compared the difference in prediction efficiency of classical ANN model, where MLP outperforms DAN2 and GARCH-MLP [4].

China's stock market has received more and more attention since this century with more scholars try to predict the fluctuation of the stocks traded on Shanghai stock exchange and Hongkong stock exchange. Moreover, since China's financial market faces a more complex global economic and political situation, any attempt to analyze the stock price movement using a linear regression model would face a dataset with high level of noise interruption. Cao et al. has proved ANN has a better performance than Fama-French's model based on the empirical data from Chinese market [5]. On account of complex regulation in IPO for mainland China financial market, it has become a praxis for Chinese high-tech companies to go public offering in Hongkong first, making the data more valuable for valuable for the research in development of China's high-tech industry. Tsang et al. argued that NN5 model is efficient in developing a buying/selling system in Hongkong stock market [6]. Nevertheless, few of the previous researches have investigated the prediction of China's internet industry, the most boosting industry in the past decade and the important driving force for economic development, making today's research valuable.

The topic of HK internet industry corporation value study is rather meaningful at the current situation. Although Chinese internet tech corporation has experienced a sharp growth contemporarily, Hongkong stock markets have seen a great decline in the first season of 2022, among which the internet industry suffers most. The reasons for the shock include but not limited to political condition, global economy and companies' financial performance etc. At this moment of panic, investors are eager to know whether they should restore faith in the internet industry or hold short position. By utilizing technical analysis based on machine learning, an explanation can be given to the rationality of price movement and even predict the future. The rest part of the paper is organized as follows. The Sec. II A gives a brief view of processing on researching data, while B describes the prediction models of the research. The Sec. III provides the results of prediction and discusses the efficiency in data explanation and flaws. Ultimately, a brief summary of the whole paper is given in the Sec. IV.

2 METHODOLOGY

2.1 Data

The study takes in less than 7 years data for total two indices (HSI and HSIII) and two stocks (Tecent and SMIC) from 17th Aug 2015 to 14th Apr 2022. Since the opening of Hang Seng Internet & Information Technology Index ("HSIII") remarks the recognition for the importance of China's internet technology corporations in Hongkong market, datasets are selected from the first day this index is invented. All the data is attained from <https://choice.eastmoney.com/>. The summary for the close price of datasets is shown in Table. 1 (taking HSI data set as an example).

This analysis used 70% of total data as the parameter selection set to carry out optimization and another 30% as the test set to compare the efficiency of model prediction. In cater to the need to solve over fitting problems in model training, parameter selection set is divided in to a validation set and a train set, each containing 35% of the total data. For FNN model, the sampling is carried out by selecting equal length samples of datasets for each year. This approach of data sampling makes each dataset more representative of the fluctuation of stock price to deal the change of data distribution from year to year, gaining a better prediction accuracy. The data summary is depicted in Table 2. As for LSTM, the sampling of the model is based on slices of time series data. The first 30% of the time series data is used as the test set, the middle 35% of the data is used as the validation set, and the last 35% of the data is used as the training set. The chronological order of the data is not disrupted in the whole process.

Table 1 The quantile and descriptive statistics summary of total data set of HSI.

Quantile statistics		Descriptive statistics	
Minimum	18319.58	**Standard deviation**	3038.23
5-th percentile	20499.16	**Coefficient of variation (CV)**	0.12
Q1	23410.48	**Kurtosis**	-0.70
median	25875.68	**Mean**	25719.50
Q3	28217.66	**Median Absolute Deviation (MAD)**	2371.54
95-th percentile	30515.59	**Skewness**	-0.13
Maximum	33154.12	**Sum**	42231421.26
Range	14834.54	**Variance**	9230845.64
Interquartile range (IQR)	4807.18	**Monotonicity**	Not monotonic

Table 2 The mean and median for each year’s train set, validation set and test set of HSI close price.

Year	**Close_trn**		**Close_val**		**Close_tst**	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
2015	21883.14	21868.39	22187.23	22140.47	22312.36	22370.04
2016	21509.66	21561.06	21383.50	21231.57	21418.01	21374.82
2017	25984.07	25808.34	26247.47	25901.10	26473.85	26849.44
2018	28767.33	28535.50	28893.12	29059.92	28901.02	29582.16
2019	27564.80	27298.28	27575.82	27614.89	27590.38	27611.47
2020	25269.16	25122.14	25303.86	24886.14	25337.81	24890.68
2021	27263.72	28050.22	27107.26	27561.34	26880.52	27723.84
2022	22928.78	23091.00	22918.08	22901.56	22730.99	22761.71

Technical indicators are useful in forecasting future stock movement. It is crucial to select the most explanatory indicators into the model to have the efficient performance. The 10 indicators this study employed follows the technical analysis system proposed by Kim and Han that has been generally accepted by some of the widely cited researches in financial machine learning field [1] (Kim [7], Kara et al. [4], Patel et al. [2]). Table 3 gives the formulas for each selected indicators and summarize the detailed statistics generated from historical data. Selected indicators and formulae are as follows:

- Simple n-day Moving Average

$$SMA_n = \frac{\sum_{i=0}^{n-1} C_{t-i}}{n} \tag{1}$$

- Weighted n-day Moving Average

$$\mathrm{WMA_n} = \frac{2\sum_{i=0}^{n-1}(n-i)C_{t-i}}{n(n+1)} \tag{2}$$

- Momentum

$$Momentum = C_t - C_{t-(n-1)} \tag{3}$$

- Stochastic K line

$$KDJK = \frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}} \times 100\% \tag{4}$$

- Stochastic D line

$$KDJD = \frac{\sum_{i=0}^{n-1} K_{t-i}}{n} \tag{5}$$

- Relative Strength Index (RSI)

$$RSI = \left(1 - \frac{\sum_{i=0}^{n-1}\frac{DW_{t-i}}{n}}{\sum_{i=0}^{n-1}\frac{DW_{t-i}}{n} + \sum_{i=0}^{n-1}\frac{UP_{t-i}}{n}}\right) \times 100\% \tag{6}$$

- Moving Average Convergence Divergence (MACD)

$$MACD(n)_{t-1} + \frac{2}{n-1} \times (DIFF_t - MACD(n)_{t-1}) \tag{7}$$

$$DIFF_t = EMA(12)_t - EMA(26)_t \tag{8}$$

$$EMA(K)_t = EMA(k)_{t-1} + \alpha \times (C_t - EMA(K)_{t-1}) \tag{9}$$

$$\alpha = \frac{2}{k+1} \tag{10}$$

- Larry William's R%

$$WR = \frac{H_n - C_t}{H_n - L_n} \times 100\% \tag{11}$$

- A/D (Accumulation/Distribution) Oscillator

$$AD = \frac{H_t - C_{t-1}}{H_t - L_t} \tag{12}$$

- CCI (Commodity Channel Index)

$$CCI = \frac{M_t - SM_t}{0.015 D_t} \tag{13}$$

$$M_t = \frac{H_t + L_t + C_t}{3} \tag{14}$$

$$SM_t = \frac{\sum_{i=0}^{n-1} M_{t-i}}{n} \tag{15}$$

$$D_t = \frac{\sum_{i=0}^{n-1} |M_{t-i} - SM_t|}{n} \tag{16}$$

Here, C_t is the closing price, L_t the low price, H_t the high price at time t, $DIFF_t = EMA(12)_t - EMA(26)_t$, EMA exponential moving average, $EMA(K)_t = EMA(k)_{t-1} + 2(C_t - EMA(K)_{t-1})/(k - 1)$, k is time period of k day exponential moving average, LL_t and HH_t mean lowest low and highest high in the last t days, respectively.

Table 3 Summary statistics for all 10 indicators (taking HSI as an example).

	Min	**Max**	**Mean**	**Std**
SMA	32779.67	18924.34	25739.69	2998.73
WMA	32807.40	-4476.05	25739.24	3003.75
Momentum	3269.89	53.55	-2.96	924.44
STC K%	100.00	0.00	53.55	32.29
STC D%	99.20	1.51	53.52	27.18
MACD	93.22	-1326.05	1.31	315.27
RSI	855.70	11.00	51.50	14.66
WILLR%	0.00	-45.59	-45.59	32.26
A/D Osc%	1.57E+08	-2.39E+07	-2.39E+07	4.88E+07
CCI	255.90	-314.32	-314.32	103.74

2.2 Prediction models

2.2.1 Feedforward Neural Networks (FNN)

As the simplest neural network model, FNN is less time consuming to construct and face fewer statistical problems like gradient explosion and gradient vanishing, but generally produce a less accurate result in prediction [8]. In this study, a four-layer FNN model is employed, taking in 10 technical indicators as neurons in the input layers and generate a single neuron in the output layer. The structure of FNN model is depicted in Fig. 1. To determine the best parameter set, w 10 potential number of neurons in the 2 hidden layers (hl1 and hl2) and 10 potential value of epoch times (ep) is introduced. The first step of parameter selection is carried out with learning rate (lr) fixed at 0.1. After the best choice of hl and ep is discovered, lr will be tested in 9 different levels in searching for the most-fit. The fluctuating parameters and potential value is given in Table 4. Since the model selection is firstly conducted on hl and ep and then lr, 9000 potential model selections are provided.

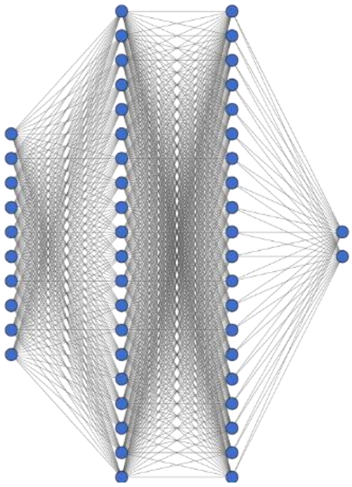


Figure 1. A sketch of FNN

Table 4 FNN parameters and their potential value tested.

Parameters	Potential values
Number of hidden layer 1/2 neurons (n)	10-100
Epochs (ep)	1000-10000
Learning rate (lr)	0.1-0.9

Parameters and methods are taken to solve the over-fitting and underfitting problem. For activation function, the ReLU is adopted whose convergence speed of SGD is proved by Krizhevsky et al. to be faster than Sigmoid/Tanh [9], whereas model's criterion and optimizer are selected as cross entropy and SGD, typical for classifying problems.

Apart from cross validation, two other methods are also introduced to solve the Underfitting and overfitting problems, namely dropout method, batch normalization for layers. According to Krizhevsky et al. [9], for the sake of avoiding overfitting, dropout method can be applied. Since one only have 1642 samples in dataset but 10 indexes for this paper, dropout method is introduced in modeling.

Layer Normalization (LN), which normalizes all neurons in an intermediate layer, is introduced to prevent over-fitting and enhance prediction accuracy, including normalizing the input data before modeling and normalizing data output from each layer before entering the activation function.

In general, a network structure that contains a large number of hidden layers will face the problem of covariate shift due to the constantly changing of parameters during training process, which will lead to underfitting of the model. According to Ioffe and Szegedy [10], this problem can be delt with batch normalization (BN). BN is to use a certain normalization method.

2.2.2 Long Short-Term Memory (LSTM)

Comparing with FNN, RNN is more suitable for time series dataset analysis. In FNN, data is processed independently as datapoint, without necessary connection with previous data. In this experiment, a four-layer RNN model is constructed to predict the index/stock price movement, with the input data extend from 10 technical indexes to additional n period previous movements of the close price. The number of backtracking periods is obtained by best parameter selection experiment. The experiment also includes the selection of best number of hidden layers, best epoch for training and best learn rate with the same process in FNN modeling.

The RNN models face the underfitting and overfitting problems just as the FNN models, so cross validation, dropout method and batch normalization for layers are also introduced. What’s more, since the memory process of RNN tend to learn the feature from the close period while the early period’s influence keeps decreasing, the LSTM is implemented to improve the memory capability of model when dealing with long-period dataset. First introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [11], LSTM model is equipped with forget gate on each layer to select useful information while forgetting the useless. This additional path of information transfer ensures the features from early data being learnt effectively.

The structure of LSTM cells is depicted in Fig2. Before model construction, Augmented Dickey-Fuller (ADF) test is introduced to examine the autocorrelation and partial autocorrelation in datasets. As a result, the number of time steps in each sample is determined to be10. Similar with the operation with FNN model, to determine the best parameter set, 10 potential number of neurons in the hidden layer (hl) and 10 potential value of epoch times (ep) are introduced. The first step of parameter selection is carried out with learning rate (lr) fixed at 0.1. After the best choice of hl and ep is discovered, lr will be tested in 9 different levels in searching for the most-fit. The fluctuating parameters and potential value the same as given in Table. 4. Since the model selection is firstly conducted on hl and ep and then lr, 9000 potential model selections are provided.

Due to observation, the distribution of factor datasets is relatively scattered, and the value ranges between factors are quite different. In order to ensure that the factors have a similar level of impact on the prediction, the factors are normalized by MinMaxScaler before input.

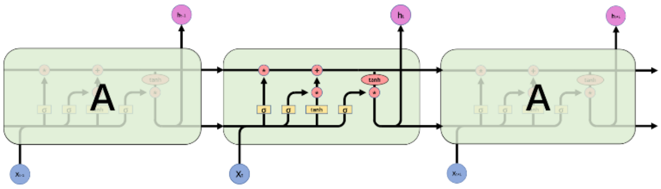
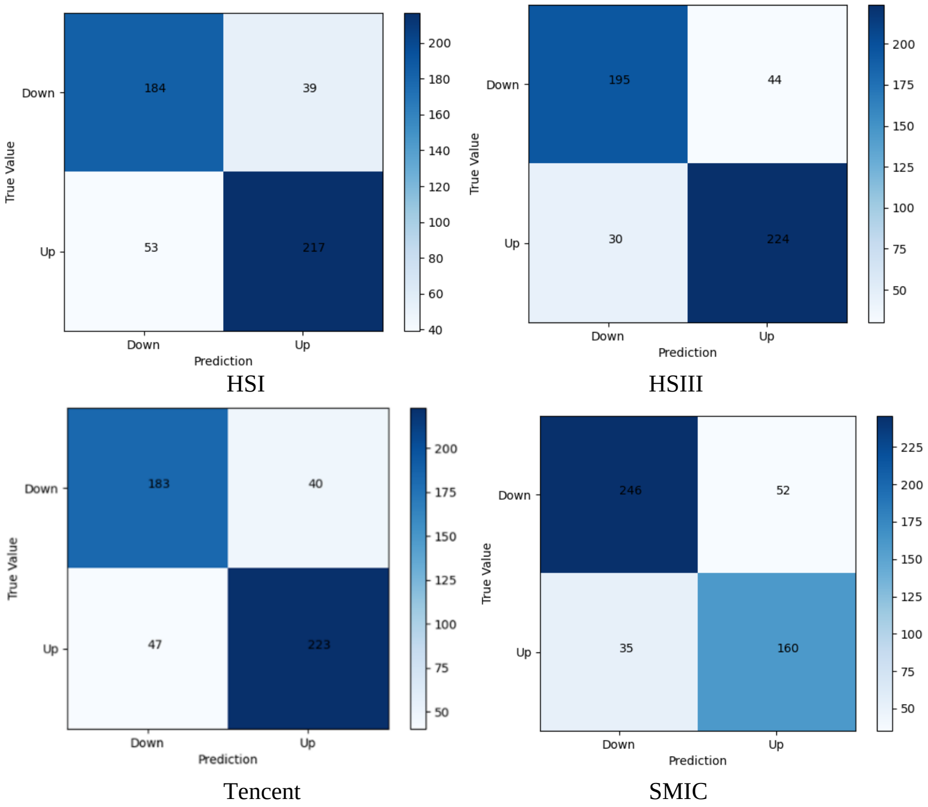


Figure 2. A sketch of LSTM Cells

3 RESULTS & DISCUSSION

Accuracy and F1-score are introduced to analyze the performance of model prediction. Accuracy is the most straightforward metric to measure a classification model, but its flaws are also obvious. For dataset whose target classification distribution is skewed, a model which is unsensitive to the certain kind of classification will also show a high accuracy. Therefore, F1-screre is introduced.



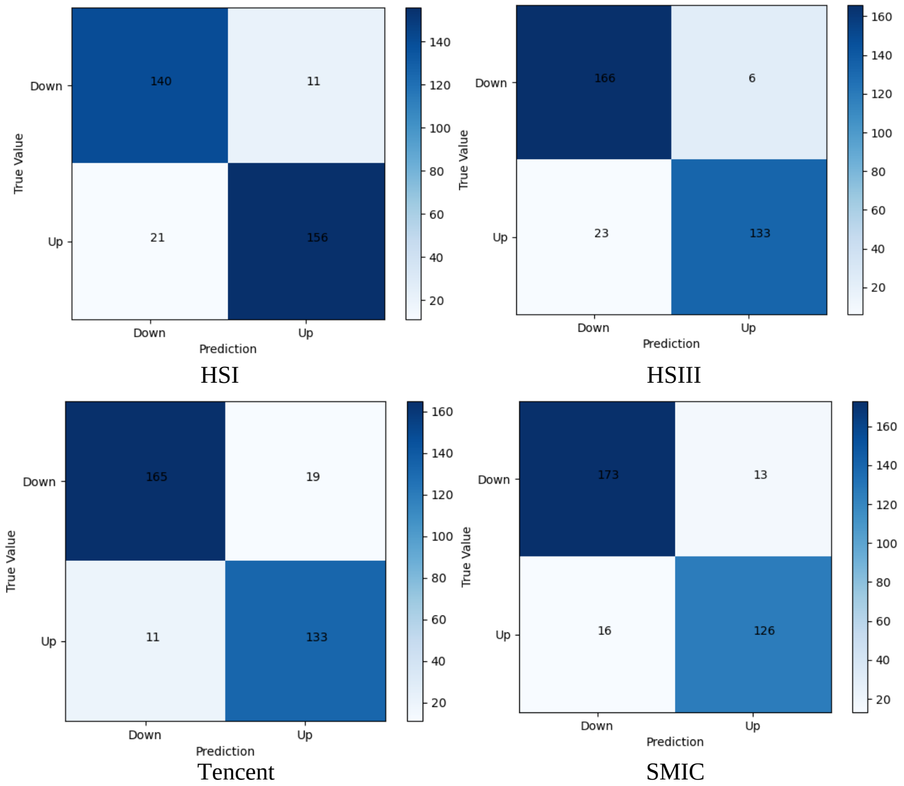
HSI

HSIII

Tencent

SMIC

Figure 3. Confusion Metrics for FNN Models



HSI

HSIII

Tencent

SMIC

Figure 4. Confusion Metrics for LSTM Models

Table 5 Comparison of models prediction for 4 datasets and their average.

Index	**FNN**		**LSTM**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
HSI	81.34%	82.51%	90.24%	90.70%
HSIII	84.99%	85.82%	91.16%	90.17%
Tencent	82.35%	83.68%	90.85%	89.86%
SMIC	82.35%	78.62%	91.16%	89.68%
Average	82.76%	82.66%	90.85%	90.10%

Table 5 gives the statistical details of FNN and LSTM prediction performance obtained from the best parameter combination for 4 datasets while Fig. 3 and Fig. 4 depict their confusion metrics, respectively. It can be learnt that FNN model using dropout and batch normalization methods is generally robust enough in stock index and individual stock prediction with both Accuracy and F1-Score reaching a level of 83%, which is almost 10% higher than prediction performance of FNN model for Indian stock and stock index without (an average accuracy of 74.94% and F1-Score of 76.59% reached) [2]. However, it is not good enough compared with LSTM prediction, which reaches an average accuracy of 90.85% and fi-score of 90.10%. The reason is that financial time series data has strong autocorrelation. On this basis, inputting the hidden state of long short-term memory to assist training will achieve better training results.

Nevertheless, this study also has flaws and shortcomings. First of all, this study focuses on the interpretation and analysis current data rather than prediction of future data. In addition, this study adopts the former's usual choice when selecting factors, and does not use widely collected and commonly used technical analysis indicators for principal component analysis (PCA) to screen out the factors with the strongest explanatory power. Last but not least, the data in this study has certain instability, and the distribution functions of the training set and the test set may be different when sampling for the LSTM model. Future research can complement these three aspects.

4 CONCLUSION

In summary, this paper investigates the stock price and index movement of Hongkong Internet technology industry based on machine learning method of FNN and LSTM. Attributed to the recent volatility in the specific market, the interpretation of prices and indices has become very valuable. According to the analysis, the machine learning method is very strong in explaining the price changes of the Hong Kong stock Internet market. The accuracy of FNN and LSTM methods for predicting the trend of price and index changes has reached more than 80%. Compared with the previous results of forecasting regional stock market markets using similar methods, the forecasting accuracy of this study has been significantly improved. At the same time, this study proves that the LSTM model has a better ability in explaining the price changes of the Hong Kong Internet industry stock market, which provides a reference for future research. Nevertheless, this study still has deficiencies in the prediction of future data and factor selection. In the future, relative researches can be carried out in these fields for a better explanation. Overall, these results offer a guideline for the comprehension modeling and explanation of Hongkong Internet technology industry stock price and index movement.

REFERENCES

[1] K. Kim, and I. Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index." Expert systems with Applications vol. 19.2, 2000, pp. 125-132.

[2] J. Patel, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert systems with applications vol. 42.1, 2015, pp. 259-268.

[3] M. Vijh, et al. "Stock closing price prediction using machine learning techniques." Procedia computer science vol. 167, 2020, pp. 599-606..

[4] Y. Kara, M. A. Boyacioglu, and Ö. K. Baykan. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert systems with Applications vol. 38.5, 2011, pp. 5311-5319.

[5] Q. Cao, M. E. Parry, and K. B. Leggio. "The three-factor model and artificial neural networks: predicting stock price movement in China." Annals of Operations Research vol. 185.1, 2011, pp. 25-44.

[6] P. M. Tsang, et al. "Design and implementation of NN5 for Hong Kong stock price forecasting." Engineering Applications of Artificial Intelligence vol. 20.4, 2007, pp. 453-461.

[7] K. Kim, "Financial time series forecasting using support vector machines." Neurocomputing vol. 55.1-2, 2003, pp. 307-319.

[8] A. Urso, et al. "Data mining: Classification and prediction." Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics 384, 2018.

[9] A. Krizhevsky, et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems, 25, 2012.

[10] S. Ioffe, and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. PMLR, 2015.

[11] S. Hochreiter, and J. Schmidhuber. "Long short-term memory." Neural computation vol. 9.8, 1997, pp. 1735-1780.