Stock Price Forecasting by Hybrid Machine Learning Techniques

Tsai, C.-F. and Wang, S.-P.

***Abstract*—Stock investment has become an important investment activity in Taiwan. However, investors usually get loss because of unclear investment objective and blind investment. Therefore, to create a good investment decision support system to assist investors in making good decisions has become an important research problem. Artificial Neural Networks (ANN) can provide relatively good performances in forecasting stock price but it cannot explain the forecasting rules clearly. On the other hand, a decision tree (DT) model can generate some rules to describe the forecasting decisions. This paper focuses on combining ANN and decision trees to create a stock price forecasting model. The experimental result shows that the combined DT+ANN model has 77% accuracy, which is higher than the single ANN and DT models over the electronic industry.**

***Index Terms*—data mining, hybrid machine learning, stock price forecasting**

I. Introduction

Stock investment is one major investment activity. However, if investors lack of enough information and knowledge, it may cause some certain loss of their investment. If we could predict stock price more accurately, we can make the society's resource allocate to a right place that avoids wasting national resource so stock market will expand healthy and people will invest more confidently to avoid blind investment behaviors.

Artificial Neural Network (ANN) is the most commonly used technique in many prediction problems. ANN has developed for many years, and it has been confirmed to provide good performances on forecasting stock price [11], [13]. However, the single usage of ANN does not allow people to understand the "decision rule" inside ANN. On the other hand, Decision trees (DT), another data mining technique is also used for forecasting stock market [9]. In particular, DT has excellent ability to describe cause and effect relation of information.

In literature, combing multiple techniques, such as classifier ensembles and hybrid techniques, has shown better performances than single techniques [11]. Therefore, the aim of this paper is to create a model which combines ANN and DT to enhance the rate of prediction accuracy in stock price forecasting as well as to provide useful decision rules.

This paper is organized as follows. Section 2 reviews related literature. Section 3 describes the research methodology. Section 4 presents the experimental results. Conclusion is provided in Section 5.

Manuscript received December 4, 2008.

Tsai, C.-F. is with the Department of Information Management, National Central University, Taiwan (phone: +886-3-422-7151; fax: +886-3-4254604; e-mail: actcft@ccu.edu.tw; cftsai@mgt.ncu.edu.tw).

Wang, S.-F. is with the Department of Accounting and Information Technology, National Chung Cheng University, Taiwan (e-mail: 92306015@nccu.edu.tw).

II. Literature Review

A. Indicators of Stock price forecasting

In literature, there are two important indicators for forecasting stock price. They are fundamental analysis, which uses the information in company's financial statement, and technical analysis, which believes that researching the trend in stock market will acquire the change rules of stock. Both of them have been used to analyze stock market [2], [14].

However, many factors, such as liner suppose exist, investor's unreasonable behavior, and the chosen data will cause some bias in fundamental analysis. In addition, as technical analysis is based on the past trend in stock market to forecast stock price, there is no evidence to display stock market having regular rules. Thus, we will combine the two analysis methods for our forecast system in order to obtain better prediction performances.

B. Artificial Neural Networks

An Artificial Neural Network (ANN) is a technique that is developed by simulating the biological nervous systems, such as the human brain. It has an excellent ability to forecast from large databases [1]. In general, ANN used to forecast stock market is based on the back-propagation algorithm.

A multilayer perceptron (MLP) neural network is used developed by the back-propagation algorithm. It consists of an input layer including a set of sensory nodes as input nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes.

C. Decision Trees

A decision tree (DT) is one of the well-known data mining techniques for many prediction problems. It also can provide reasonable classification and forecasting performances. DT has a tree structure that depends on different situations to create a number of nodes and branches. Every node presents an output/target class and every branch presents a process/decision for classification. The end node presents a forecasting result. After the decision tree model is established, we can compute the error rate to pruning the decision tree. Pruning is a behavior to improve the forecasting ability and classification ability of decision trees

and make decisions more efficiently. Different from ANN, DT generates a number of decision rules for further analyses.

D. Related Work

Table 1 compares related work of stock price forecasting in terms of their indicators used, models created, and findings.

Table 1 Comparisons of related work

Work	Forecasting model	Indicator	Finding
Phua et al. (2003) [12]	ANN	Technique indexes	ANN has average success rate above 60% and the best prediction result is 74 %
Chen et al. (2003) [1]	PNN GMM	Technique indexes	PNN obtain higher return than other investment strategy and GMM
Chang et al. (2004) [11]	Back-propagation ANN Logic ARIMA	Fundamental indexes Technique indexes	Hybrid model has better forecast ability than single model
O'Connor and Madden (2006) [10]	ANN	Fundamental indexes	ANN has forecast ability in stock market because it has better return than overall stock market
Roh (2006) [4]	ANN, NN-EWMA, NN-GARCH and NN-EGARCH	Macroeconomic indexes	Hybrid ANN has better forecast ability than single.
Kunhuang and Tiffany (2006) [7]	Back-propagation ANN and Chen's model (time series model)	Technique indexes	ANN has better forecast ability than time series model
Zhu et al. (2007) [17]	ANN	Technique indexes	ANN can forecast stock index increment and trading volume will lead to modest improvements in stock index performance
Wang (2007) [16]	Hybrid ANN	Technique indexes	ANN combined other techniques has good forecast ability in stock market
Hassan et al. (2007) [8]	ANN HMM GA	Technique indexes	The fusion model has better forecast ability than single and it has good forecast ability as good as ARIMA model
Kim and Shin (2007) [6]	ANN ATNNs TDNNs GAs	Technique indexes	Hybrid model has better forecast ability than single and ANN has ability to forecast stock market

Regarding Table 1, we find that ANN has been used widely and can provide better forecasting ability on stock market than other models, especially in nonlinear questions. However, ANN can be regarded as a black box system that it can not describe the cause and effect. Therefore, we propose to combine ANN and decision trees to show the cause and effect in ANN and maintain the excellent forecasting ability of ANN.

III. Research Methodology

A. Dataset

This experiment uses fundamental analysis and technical analysis as the indicators for ANN and decision trees to forecast the stock price in electron industry in Taiwan, which is the major industry of the stock market investment in Taiwan. We choose fundamental indexes, technical indexes, macroeconomic indexes, and the dataset comes from the TEJ database.

After all the related variables are collected, the collected data are pre-processed by principle component analysis (PCA) for filtering out unrepresentative variables. In particular, variables that have mutuality above 0.9 are selected. As a result, there are 53 variables selected. Then, the pre-processed data will be divided into training and testing data respectively to train and test the forecasting models which are ANN, decision trees, ANN combined with decision trees (ANN+DT), and decision tree combined with decision tree (DT+DT).

On the other hand, for obtaining the optimal forecasting performance, we use the sliding windows method to find the best parameter of developing the ANN forecasting model shown in Fig. 1. Eighty percent data are used for training and the remainder is used for testing. Therefore, the model is trained and tested four times. Then, the four different results are averaged for the forecasting performance of the model.

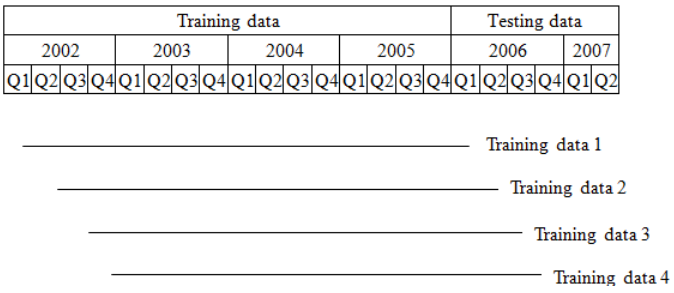

Figure 1 The sliding window

B. Prediction Models - The ANN model

We use the back-propagation ANN and Delta-Rules method to establish the ANN model. Delta-Rules method means when the ANN learns a new data the weight average in the network will update every time. The parameters of creating the ANN model are described below.

- Hidden layer: Related work, such as [5], has found that using one hidden layer of ANN can provide better performances. Therefore, we consider one hidden layer to create the ANN model to forecast stock price.
- Hidden layer node: In literature, there is no exact answer to the number of the hidden layer nodes. In particular, the over-fitting problem will occur if too many nodes are considered, and vice versa. Therefore, it is problem dependent and the try and error and cross validation methods will be used to find the optimal setting of the ANN model. In this paper, we consider 8, 16, 24, and 32 as the number of hidden layer nodes.
- Training Epoch: Similar to the hidden layer nodes,

over-training could occur if the training epoch is not well set. Therefore, 50, 100, 200, and 300 training epochs are considered.

C. *Prediction Models – Decision Trees*

The CART (classification and regression tree) technique is used to create the decision tree model. The output result of the decision tree model will be 1 and 0, in which 1 means stock price will rise and 0 means stock price will fall. The default value is used to establish the initial decision tree and then pruning the least related variables to find the best model.

D. *Model Combinations*

At first, when the ANN and DT models are constructed individually as the baselines, the test data is used to test these two models in order to combine the second level prediction model. Then, we use the following steps shown in Fig. 2 to train and combine ANN with DT (ANN+DT) as a hybrid model.

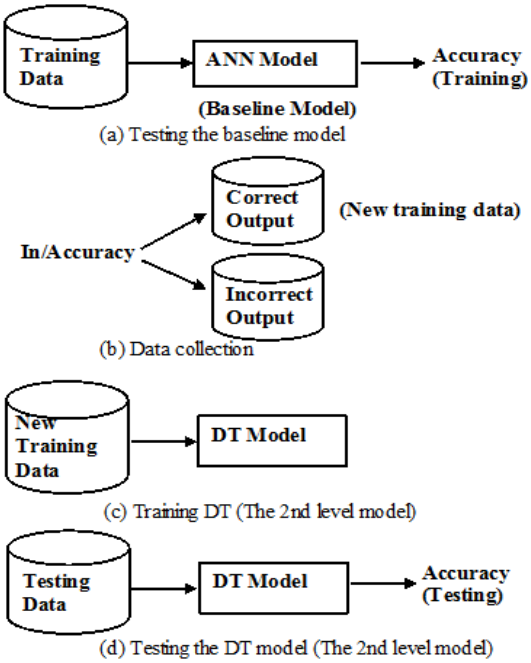

Figure 2 The process of combining ANN with DT

In addition, we also combine two decision trees (DT+DT) for further comparisons. Therefore, the process to combine two DT models is shown in Fig. 3.

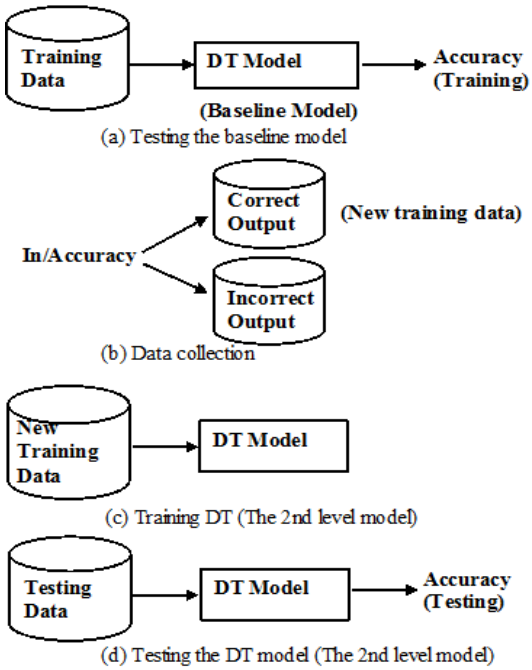

Figure 3 The process of combining two DT models

E. *Evaluation Methods*

To evaluate the performance of the baselines and combined models, in this paper we consider average prediction accuracy and the Type I & Type II errors. The Type I & Type II errors can be calculated by a confusion matrix shown in Table 2.

Table 2 Confusion matrix

		Predict	
		Fall	Rise
Actual	Fall		I
	Rise	II	

The Type I error means that the output of the prediction model is 'rise', but the stock actually falls. On the other hand, the Type II error means that the output of the prediction model is 'fall', but the stock actually rises.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This paper compares four methods including two baselines, which are ANN and DT and two combined models, which are ANN+DT and DT+DT in order to evaluate their prediction performances.

To construct these four models, four different training datasets and five different testing datasets to create our baseline models over the electronic industry. The duration of the training data and testing data are shown in Table 3.

Table 3 The training and testing datasets

	Duration
training data1	02Q1-06Q1
training data2	02Q2-06Q2
training data3	02Q3-06Q2
training data4	02Q4-06Q3
testing data1	06Q1
testing data2	06Q2
testing data3	06Q3-06Q4
testing data4	06Q3-07Q1
testing data5	06Q3-07Q2

Note that the testing data includes Q1-Q4 to prove that the model can be used in every season. The training data 3 and 4 have a shorter period because we want to avoid the training data cover the testing data too much.

A. *ANN*

For the baseline ANN model, after testing a number of different learning epochs and hidden layer nodes, we find that the best learning epoch is 500 and the hidden layer nodes is 10. Table 4 and 5 show the prediction performances of ANN in terms of prediction accuracy and Type I & II errors respectively.

On average, ANN provides 59.016% prediction accuracy and 49.5% and 43.7% for Type I and II errors respectively.

Table 4 Prediction accuracy of ANN

	test data1	test data2	test data3	test data4	test data5	average	model average
ANN1	71.74	24.64	15.58	22.95	18.84	30.75	59.016
ANN2	71.74	75.36	84.06	76.81	59.24	73.442	
ANN3	71.74	74.64	81.88	75.36	58.33	72.39	
ANN4	71.74	74.64	52.17	55.56	43.3	59.482	

Table 5 Type I & II errors of ANN

		Predict	
		Fall	Rise
Actual	Fall	196	**192**
	Rise	**494**	636

B. *Decision Trees*

Table 6 and 7 show prediction accuracy and Type I & II errors of DT respectively.

Table 6 Prediction accuracy of DT

	test data1	test data2	test data3	test data4	test data5	average	model average
DT1	71.14	25.36	72.83	69.32	69.93	61.716	65.415
DT2	71.74	24.64	84.06	76.33	80.62	67.478	
DT3	71.74	24.64	79.71	73.19	76.81	65.218	
DT4	27.54	78.99	79.71	73.19	76.81	67.248	

Table 7 Type I & II errors of DT

		Predict	
		Fall	Rise
Actual	Fall	50	**338**
	Rise	**242**	888

On average, the decision tree model provides 65.415% prediction accuracy and 87.1% and 21.4% for Type I and II errors respectively. This is interesting that although this model performs better than ANN for the rate of prediction accuracy and type II errors, it performs poorly for the Type I error.

C. *Combining DT with ANN*

After developing the two different baseline models, we find that the decision tree model provides better performances than ANN. This implies that DT is likely to extract more noisy data or outliers than ANN for the combined hybrid model. Therefore, we decide to use the decision tree baseline as the first model to filter out noisy data or outliers for the second model based on ANN. This is opposite to the process shown in Fig. 2. As a result, the DT + ANN hybrid model is constructed. Table 8 and 9 show the performances of the combined model.

The combined DT + ANN model provides 77.1985% prediction accuracy and 71.9% and 1.95% for the type I & II errors. Similar to the decision tree model, it performs better for the type II errors, but the type II error is not satisfactory.

Table 8 Prediction accuracy of DT +ANN

	test data1	test data2	test data3	test data4	test data5	average	model average
DT1+ANN	75.36	40.58	82.25	75.6	79.17	70.592	77.1985
DT2+ANN	73.19	84.06	85.14	76.81	80.98	80.036	
DT3+ANN	77.54	84.06	84.06	76.09	79.89	80.328	
DT4+ANN	77.54	75.36	81.88	74.88	79.53	77.838	

Table 9 Type I & II errors of DT + ANN

		Predict	
		Fall	Rise
Actual	Fall	109	**279**
	Rise	**22**	1108

D. Combining Two DT Models

On the other hand, the alternative is to combine two decision trees model, i.e. DT + DT. The prediction performances are shown in Table 10 and 11 for prediction accuracy and Type I & II errors respectively

Table 10 Prediction accuracy of DT + DT

	test data1	test data2	test data3	test data4	test data5	average	model average
DT+DT1	71.74	38.41	73.91	70.05	70.83	64.988	66.8515
DT+DT2	71.74	24.64	84.06	76.81	80.68	67.586	
DT+DT3	77.54	28.26	80.07	73.67	77.54	67.416	
DT+DT4	77.54	28.26	80.07	73.67	77.54	67.416	

Table 11 Type I & II errors of DT + DT

		Predict	
		Fall	Rise
Actual	Fall	45	**343**
	Rise	**75**	1055

On average, the DT+DT model provides 66.8515% accuracy and 88.4% and 6.64% for The type I & II errors.

E. Comparisons and Discussions

For the comparison between these four models shown in Table 12 and Fig. 4, we can discover that the DT+ANN model has better forecasting performances than the others. On the other hand, Table 12 compares the type I & II errors of these four models.

As we can see that the DT + ANN model performs the best over the average prediction accuracy. One reason is that because it provides the lowest rate of type II errors, 1.95%, which can result in higher average prediction accuracy. Therefore, this model is particular useful to predict rising stocks. On the other hand, all of these four models do not provide relatively low Type I error rates to predict falling stocks.

Fig. 5 shows the decision rules of the DT followed by the DT + ANN, which can be used to describe the forecasting decisions. That is, the correct outputs of the second ANN model are used to train a DT model.

Table 12 Comparisons of prediction accuracy

	test data1	test data2	test data3	test data4	test data5	average	model average
DT1	71.14	25.36	72.83	69.32	69.93	61.716	65.415
DT2	71.74	24.64	84.06	76.33	80.62	67.478	
DT3	71.74	24.64	79.71	73.19	76.81	65.218	
DT4	27.54	78.99	79.71	73.19	76.81	67.248	
ANN1	71.74	24.64	15.58	22.95	18.84	30.75	59.016
ANN2	71.74	75.36	84.06	76.81	59.24	73.442	
ANN3	71.74	74.64	81.88	75.36	58.33	72.39	
ANN4	71.74	74.64	52.17	55.56	43.3	59.482	
DT1+ANN	75.36	40.58	82.25	75.6	79.17	70.592	77.1985
DT2+ANN	73.19	84.06	85.14	76.81	80.98	80.036	
DT3+ANN	77.54	84.06	84.06	76.09	79.89	80.328	
DT4+ANN	77.54	75.36	81.88	74.88	79.53	77.838	
DT+DT1	71.74	38.41	73.91	70.05	70.83	64.988	66.8515
DT+DT2	71.74	24.64	84.06	76.81	80.68	67.586	
DT+DT3	77.54	28.26	80.07	73.67	77.54	67.416	
DT+DT4	77.54	28.26	80.07	73.67	77.54	67.416	

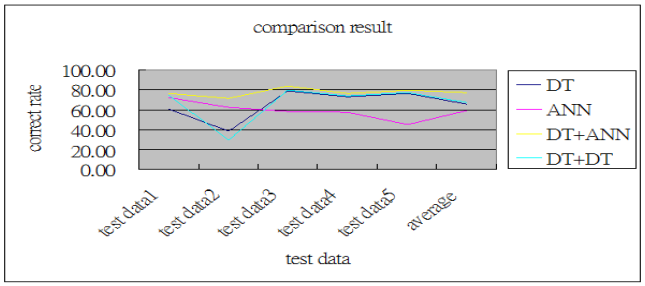

Figure 4 Prediction accuracy of the four models

Table 13 Type I & II errors of the four models

	Type I error	Type II error
DT	87.1%	21.4%
ANN	49.5%	43.7%
DT + ANN	71.9%	1.95%
DT + DT	88.4%	6.64%

V. Conclusion and Future Work

Stock investment has become an important fashion in Taiwan. However, as the economic environment is changed so fast and the investment objectives are vary, investors are difficult to make correct decisions for stock investment because they lack of a detached systemic decision support tools. Therefore, this paper combines the commonly used ANN techniques and the great explanation ability of decision tree for a better decision support system in order to help investors to make more correct decision in stock investment.

By developing the ANN and DT baseline models, two combined models are also compared, which are DT + ANN and DT + DT. The sliding window method was used to examine the performance of these models. Regarding the experimental result, the DT + ANN hybrid model has 77%

forecasting accuracy in electronic industry stock and supply reliable rules to assist investors to determine when to perform their investment. In particular, there are twelve different decision rules for predicting the stock price rise and fall.

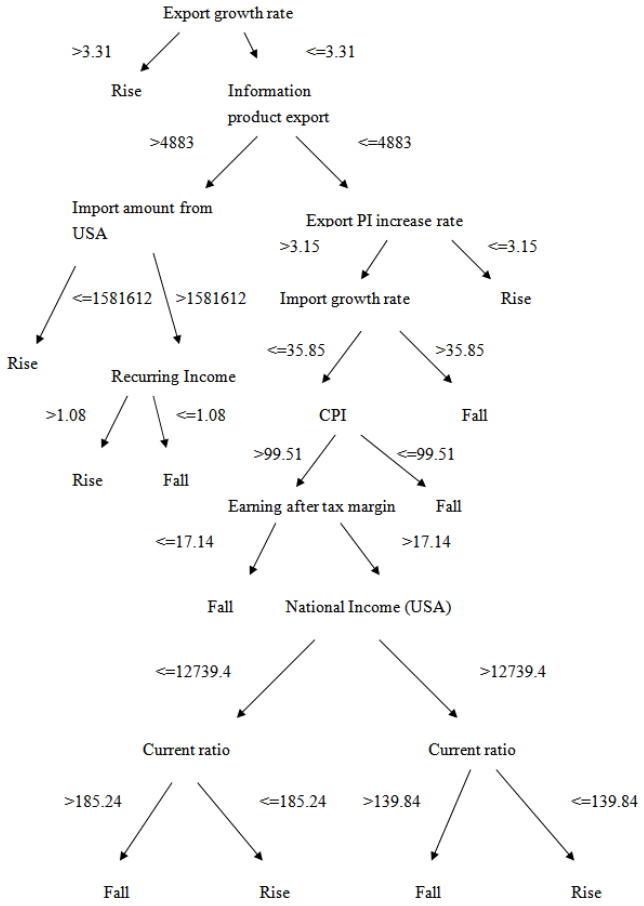

Figure 5 The decision tree structure

For future work, several issues could be considered. First, other techniques, such as support vector machines, genetic algorithm, etc. can be applied for further comparisons. Second, dimensionality reduction using other advanced methods, such as genetic algorithm can be used to pre-process the data in order to obtain better prediction performances. Finally, other industries in addition to the electronic one can be further considered for comparisons.

REFERENCES

[1] A. S. Chen, M. T. Leung, and H. Daouk, " Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index" ,*Computers & Operations Research*, 2003, vol. 30, pp. 901-923.

[2] D. Olson and C. Mossman, "Neural network forecasts of Canadian stock returns using accounting ratios", *International Journal of Forecasting*, 2003, vol. 19(3), pp. 453-465.

[3] G. Ramazan, "Non-linear prediction of security returns with moving average rules", *Journal of Forecasting*, 1996, vol. 15(3), pp. 165-174.

[4] T. H. Roh, "Forecasting the volatility of stock price index", *Expert Systems with Applications: An International Journal*, November 2007, vol. 33, pp. 916-922.

[5] G. S. Swales and Y. Yoon, "Applying artificial neural network to investment analysis", *Financial Analysts Journal*, 1992, pp. 78-80.

[6] H. J. Kim, and K. S. Shin, "A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets", *Applied Soft Computing*, 2007, vol. 7, pp. 569–576.

[7] H. Kunhuang, and T. H. K. Yu, "The application of neural networks to forecast fuzzy time series", *Physical A: Statistical Mechanics and its Applications*, 2006, vol. 363(2), pp. 481-491.

[8] H. M. Hassan, A. K. Hamadi, and S. M. Saleem, "Towards Evaluation of Phonics Method for Teaching of Reading Using Artificial Neural Networks (A Cognitive Modeling Approach)", *IEEE International Symposium on Signal Processing and Information Technology*, 2007, pp. 855-862.

[9] J. L. Wang and S. H. Chan, "Stock market trading rule discovery using two-layer bias decision tree", *Expert Systems with Applications*, 2006, vol. 30, pp. 605–611.

[10] N. O'Connor and M. G. Madden, "A neural network approach to predicting stock exchange movements using external factors", *Applications and innovations in intelligent* network to investment analysis", *Financial Analysts Journal*, 1992, pp. 78-80.

[11] P. C. Chang, Y. W. Wang, and W. N. Yang, "An investigation of the hybrid forecasting models for stock price variation in Taiwan.", *Journal of the Chinese Institute of Industrial Engineering*, 2004, vol. 21(4), pp. 358–368.

[12] P. K. H. Phua, X. T. Zhu, and H. K. Chung, "Forecasting Stock Index Increments Using Neural Networks with Trust Region Methods", *Proceedings of the International Joint Conference on Neural Networks*, 2003, vol. 1, pp. 260-265.

[13] P. R. Burrell, and B. O. Folarin, "The impact of neural networks in finance", *Neural Computing & Applications*, 1997, vol. 6, pp. 193–200.

[14] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks", *IJCNN International Joint Conference on Neural Networks*, 1990, vol. 1, pp. 1-6.

[15] T. S. Quah and B. Srinivasan, "Improving returns on stock investment through neural network selection", *Expert Systems with Applications*, 1999, vol. 17, pp. 295–301.

[16] Y. H. Wang, "Nonlinear neural network forecasting model for stock index option price: Hybrid GJR–GARCH approach", *Expert Systems with Applications*, January 2009, vol. 36(1), pp. 564-570.

[17] X. Zhu, H. Wang, L. Xu, and H. Li, "Predicting stock index increments by neural networks: The role of trading volume under different horizons", *Expert Systems with Applications*, 2007, vol. 34(4), pp. 3043-3054.