MANCHESTER 1824
The University of Manchester

The University of Manchester Research

Machine learning approach for crude oil price prediction with Artificial Neural Networks-Quantitative (ANN-Q) model

DOI:
10.1109/IJCNN.2010.5596602

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Abdullah, S. N., & Zeng, X. (2010). Machine learning approach for crude oil price prediction with Artificial Neural Networks-Quantitative (ANN-Q) model. In *Proceedings of the International Joint Conference on Neural Networks|Proc Int Jt Conf Neural Networks* IEEE. https://doi.org/10.1109/IJCNN.2010.5596602

Published in:
Proceedings of the International Joint Conference on Neural Networks|Proc Int Jt Conf Neural Networks

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:17. Feb. 2020

Machine Learning Approach for Crude Oil Price Prediction with Artificial Neural Networks-Quantitative (ANN-Q) Model

S. N. Abdullah, X. Zeng, *School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL United Kingdom*

Abstract— **The volatility of crude oil market and its chain effects to the world economy augmented the interest and fear of individuals, public and private sectors. Previous statistical and econometric techniques used for prediction, offer good results when dealing with linear data. Nevertheless, crude oil price series deal with high nonlinearity and irregular events. The continuous usage of statistical and econometric techniques for crude oil price prediction might demonstrate demotions to the prediction performance. Machine Learning and Computational Intelligence approach through combination of historical quantitative data with qualitative data from experts' view and news is a remedy proposed to predict this. This paper will discuss the first part of the research, focusing on to (i) the development of Hierarchical Conceptual (HC) model and (ii) the development of Artificial Neural Networks-Quantitative (ANN-Q) model.**

I. Introduction

THE drastic increment of global crude oil market in July 2008 had indeed affects the economy around the globe. Started with USD$69.00 per barrel in April, 2006 and rising up to 50% increment of USD$134.00 per barrel in July, 2008, this phenomenon had then gave an exclusive economic impact to the oil importing and exporting countries. In addition, crude oil products are one of the world's major commodities with high volatility level. They are traded in New York Mercantile Exchange (NYMEX) market together with other energy and mineral commodities. Simultaneously, the volatility of this crude oil price is depending on demand and supply of the commodity, level of inventories, economic indicators and finally, the population of the world.

Due to strong chain effects owned by this crude oil market, any changes in the factors involved will have exclusive impact to the price. Furthermore, the crude oil price contributes over 50% on the average price of petroleum and it is one of the most used commodities around the globe. Therefore, every increment and decrement that occurs to the crude oil price will then also give impact to the price of petroleum and later correspond to the global economy. A good prediction tool is crucial to be developed for this matter. From [1], it is proved that among the main factors affect the volatility of the crude oil price are the demand and supply of the oil, population, political climates and also economical aspects. In [1], a survey is given to a large geographical dispersed of experts to impose their common responses on future changes of oil prices and related economic variables. This forecast survey was done through series of questionnaires and intended to tailor the California Energy Commission's 20 years of planning period. Moreover, the variable used for this forecast is the average price of internationally traded oil, West Texas Intermediate (WTI). In this survey, experts are considered to forecast and rank various aspects of the crude oil price changes by High, Most likely and Low based on the factors provided in the questionnaires.

In this paper, we will discuss the research background and its literature review in section II, problems identified from earlier research in section III and the research methodology used in section IV. Later, the simulation study and its empirical result will be discussed in section V and VI. Finally, section VII will discuss on the conclusion and the possible future works for the study.

II. Research Background

Crude oil price market prediction is known for its obscurity and complexity. Due to its high vacillation degree, unpredictable irregularity events, and the complex correlations involved between the factors in the market, it is indeed difficult to predict the movements of the crude oil price. E. Panas et. al [2] mentioned that crude oil market has strong evidence of chaos and develops as one of the most volatile market in the world. Corresponding to that, there are few numbers of research conducted for crude oil price prediction. Among the research model used are single statistical and econometric model, single Artificial Intelligence (AI) model and the hybrid.

Formerly, Generalized Autoregressive Conditional Heteroskedasticity **(**GARCH) model and Naive Random Walk were among the statistical and econometric model used to predict the crude oil price. Research [3] successfully utilised a probabilistic model to predict the oil price. The research was conducted based on a case study about the probabilistic inheritance of Belief Network (BN) models.

Manuscript received February 6, 2010. This work was supported by the Malaysian Government under *Majlis Amanah Rakyat (MARA)*-Excellent Students Scheme.

S. N. Abdullah, is with the School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL United Kingdom (phone: +44 (0) 161 275 7520; fax: +44 (0) 161 275 6204; e-mail: sitinorbaiti.abdullah@postgrad.manchester.ac.uk).

X. Zeng, is with the School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL United Kingdom (phone: +44 (0) 161 276 3362; fax: +44 (0) 161 275 6204; e-mail: x.zeng@manchester.ac.uk)

978-1-4244-8126-2/10/$26.00 ©2010 Crown

The models are used to forecast crude oil price and then produce a probabilistic prediction for it [4]. The probabilistic prediction is actually generated by running Monte Carlo analyses on annual WTI average prices. For the purpose of simulation experiment in [4], the analysis done in this study is based on two assumptions of the timing when Iraq's return to the market and the impact of oil exports from the Former Soviet Union. Three variables input are then used to define the scenarios; the probabilities of embargo ends, total demand and other world productions. The result from the simulation were robust and consistent with the annual average prices are almost certain in between USD$15.00 to USD$25.00 per barrel. There was only 0.75% out of the total scenarios, predicted price over the range.

Other statistical model predictions made for crude oil price is by C. Morana [5]. This research used semi parametric approach suggested in [6] for short-term oil price prediction. It also used GARCH to employ oil price changes to predict the oil price distribution over short-term horizon. The approach used one-month-ahead daily Brent oil price which emphasised on periods with the high uncertainty (November 21, 1998 to January 21, 1999). Furthermore, the analysis of the forecasting is based on the last two months of the available data and according to the analysis the result was strayed from the actual. This is most likely linked to the widening of the forecast confidence interval. Nevertheless, the study offers improvement from [6].

Next statistical model used for predicting the crude oil market is by [7] where they predict monthly WTI spot price using relative inventories. This study used Relative Stock Model (RSTK) as the basis to predict the price by comparing two other alternative models; Naïve Autoregressive (NAIV) forecast model and Modified Alternative (MALT) model. The only variable they used in this research is the petroleum inventories because of its independent practicality and it is readily available every month. RSTK model shows the best performance for both in and out of sample forecast compared to the other two models. It is also being used by the Energy Information Administration (EIA) with among other models to investigate the future market disruptions that derived from changes in demand and production.

Nowadays, AI models are among the popular tools to be used for prediction. As an alternative tool to statistical and econometric models, AI offers recognition ability on complex patterns and also on providing intelligent reasoning and intelligent decision-making based on data. Among the single AI models used for predicting the crude oil price is Support Vector Machine (SVM) in [8] where for the task of time-series prediction, this research focused only on Support Vector Regression (SVR) model. In this study, they used monthly WTI price ranging from January, 1970 to December, 2003 as the only independent variable. There was no normalisation process involved in the investigation as to make it simple and moreover, SVM is resistant to noise. This model is evaluated with other two models; Auto-Regressive Integrated Moving Average (ARIMA) and Backpropogation Neural Networks (BPNN). The evaluation shows that SVR outperformed the other two mentioned. However, the authors do admitted that BPNN was also outperformed in some of the sub-periods evaluated and has the capability of capturing the nonlinear dynamics of crude oil price.

Although both statistical and AI models performed well in their individual approach, there are still some scarceness and limitations that can be improved for a better performance tool. Later, hybrid models are introduced to treat these nuisances. Study [9] acquainted TEI@I methodology to hybrid four models for the prediction. Based on the idea of combining Text Mining, Econometrics, Intelligence (intelligent algorithm) components and integrating (@) of the mentioned, this study integrates Web-based Text Mining (WTM), Auto-Regressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), and Rules-based Expert System (RES) to predict the price. The dynamic movements of crude oil price market are also due to related irregular events that occurred unexpectedly. Therefore, ARIMA and ANN are used to handle the linear and nonlinear components respectively in crude oil price while WTM and RES as the news or irregular events retriever. Later, these four models are integrated with Nonlinear Integrated Forecasting approach based on BPNN training where it makes the sum of squared errors minimal. By using monthly WTI price together with the online news as the training data, this approach performed very well in predicting the crude oil price.

Another hybrid models related to crude oil price forecasting is [10], a rough-set refined text mining approach where text mining and rough-set are combined to produce useful knowledge that can be used to configure and predict the tendency of crude oil market. The advantage of this approach is that it can consider both the quantitative and the qualitative factors. The model input variables used are all possible events that affecting the crude oil market. These events are extracted via internet and internal file system using the rough-set refined text mining approach. Other than that, world oil demand and supply, crude oil production and crude oil stock level are selected as the input variables and monthly WTI price as the output variable used. Nonetheless, this approach has demonstrated a promising tool for predicting the movement of crude oil market where it outperformed the other models in the evaluation process.

Finally, research [11] integrates Empirical Mode Decomposition (EMD) with Feed-forward Neural Network (FNN) and Adaptive Linear Neural Network (ALNN); EMD-FNN-ALNN to formulate an ensemble output for the original crude oil price series. This study used daily WTI and Brent oil price ranging from January, 1986 to September, 2003, excluding public holidays. From the experiment evaluation, we can conclude that this method offers an alternative prediction tool to crude oil price forecasting. It also proved that the decomposition and ensemble techniques used in EMD (Decomposition)-FNN (Prediction)-ALNN (Ensemble) had improved the limitations carried by other previous single models. Details of this approach can be referred in [11].

Next, section III we will discuss on the problems encountered from the investigation made from the previous research.

III. Problem Statement

There are five main problems identified based on investigations made on previous research. Firstly, data used in the previous predictions are majority employed from WTI or Brent crude oil price without taking into consideration other inputs that are involved together in the market. The crude oil price market volatiles from the contributions made by other factors surround it and neglecting these factors will demote the capability of a prediction tool. A good prediction is the one that can comprehends and correlates between factors, sparks information on the trend and finally, predict it accurately.

Secondly, there are scarce numbers of research that implement the verification and validation technique on the main factors involving in the fluctuation. Besides the global crude oil price, other popular factors that being used in previous research are demand and supply. Although, demand and supply of oil plays vital role to the market volatility, the use of these observations only is not enough to comprehensively render the information offered by the trend. There are also other factors that contributed to the trend and gave impact to the price. Therefore, by embracing appropriate key factors and later correlate them will help to achieve a thorough and comprehensive prediction for the market.

Thirdly, time-series data are mainly used for prediction. Nevertheless, data pre-processing and data representation process are made absent in some of the previous research. These two processes are important to cleanse and reduce errors and noises in data set and uniform it. Later, these will help to organise the process of prediction, make it more systematic and finally, generates more stable result. Without these processes, the prediction tool will be less reliable.

Fourthly, the crude oil price movement was the popular topic studied previously and not the crude oil price itself. Predicting the movement of the price only is not sufficient to characterise the market where else, crisp prediction will offer far more persona. A prediction on the movement together with the price itself will tender more usable, discrete and practical implementation to the real world problem.

Sincerely, the practicability of the previous study is still dubious as the crude oil market itself is chaotic. Still, there are opportunities for improvement in the future as the advancement of our world technology is rapid.

In section IV, we will discuss further the methodology used for our research by explaining the models involved in sub-section (*A)* and (*B).*

IV. Research Methodology

The vacillation of the crude oil market is dependable on factors that contribute to the environment. Consequently, every change in either one of the key factors will give direct impact on the market. Hence, selecting the appropriate key factors that contribute to this volatility is crucial. Based on discussion in II and III, common factors used as variables in prediction are almost similar by most of the studies. Concurrently, there are other important factors that should not be neglected as they also have their analogous influence to the market like the demand and supply. Nevertheless, understanding about the market is a vital aspect to gain before we can predict the price. One of the ways to understand the market is by verifying the appropriate information.

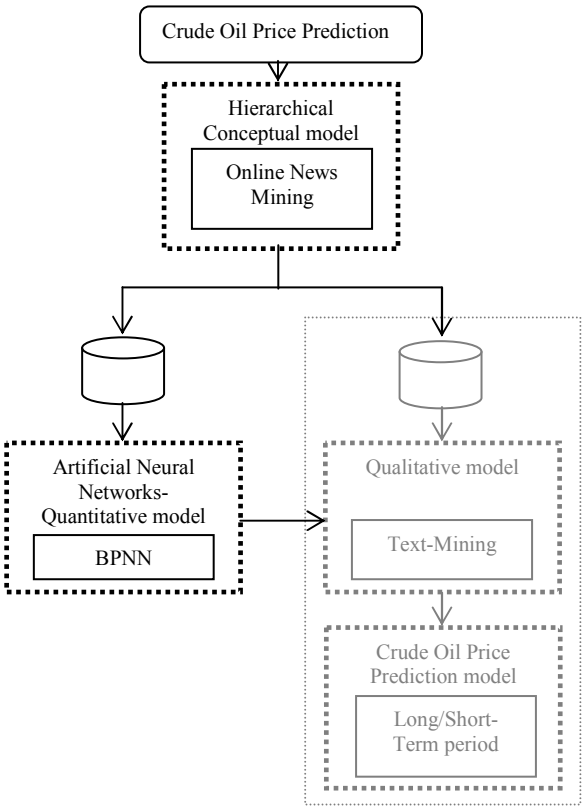

Fig. 1. Machine Learning Approach for Crude Oil Price Prediction: The Overall Research Methodology

Generally, this research use machine learning and computational intelligence approach to integrate the historical quantitative data derived from various key factors affecting the price, together with the qualitative data composed from experts' views and news to predict crude oil price for long and short term period. In this paper, we will only discuss on the first part of the methodology where we emphasised on the development of Hierarchical Conceptual model and also the Quantitative model for long-term prediction in section (*A)* and (*B)* respectively. The overall methodology for this research is as represented in Fig.1.

A. Hierarchical Conceptual (HC) Model

In developing a prediction model, factors related to the price fluctuations are first need to be verified and validated to ensure the appropriateness and the relevancy of factors to be used. Hence, Hierarchical Conceptual (HC) model is developed to fulfil this purpose. To understand the vacillation of the crude oil price, key factors that reflected to this situation are first retrieved from the online news. Online

news is a great source of information as it stores current information which contains rules that represent the key factors involved.

Simultaneously, text mining approach is applied together with Google News to mine the news online. Information regarding on the knowledge and rules of the market are then verified and stored in database according to its quantitative or qualitative features. A total of 22 numbers of quantitative factors were discovered from this model and used as the input variables compositing sub-factors of supply, demand, inventory, economy, population and inclusive of WTI price as the output variable. These quantitative key factors are presented in Table I.

TABLE I
THE QUANTITATIVE KEY FACTORS INFLUENCING CRUDE OIL PRICE

Variables	Factors
S^T	Supply
S_{a1}	Productions of OPEC countries
S_{a2}	Productions of Non-OPEC countries
S_{b1}	Proved reserves of OPEC countries
S_{b2}	Proved reserves of OECD countries
S_{c1}	Number of well drilled
D^T	Demand
D_{a1}	Consumption of OECD countries
D_{a2}	Consumption of China
D_{a3}	Consumption of India
I^T	Inventory
I_{a1}	Ending stocks of OECD countries
I_{a2}	Ending stocks of US
I_{b1}	US petroleum imports from OPEC countries
I_{b2}	US petroleum imports from Non-OPEC countries
I_{c1}	US crude oil imports from OPEC countries
I_{c2}	US crude oil imports from Non-OPEC countries
E^T	Economy
E_{a1}	Foreign Exchange of GBP/USD
E_{a2}	Foreign Exchange of Yen/USD
E_{a3}	Foreign Exchange of Euro/USD
E_{b1}	US Growth Domestic Products (GDP)
E_{c1}	US Inflation rate
E_{d1}	US Consumer Price Index (CPI)
P^T	Population
P_{a1}	Population of developed countries
P_{a2}	Population of less developed countries
WTI	West Texas Intermediate price

OPEC: Organisation of the Petroleum Exporting Countries
OECD: Organisation for Economic Co-operation and Development

There are three stages involved to retrieve the key factors; (i) data classification, (ii) information retrieval and (iii) feature extraction.

(i) Data Classification

Data from monthly WTI price are first being classified into three classes. This classification is made based on the turning points that are represented by the price trend. The three classes are; (a) Large Impact class (more than 30% variance between turning points), (b) Medium Impact class (20-30% variance) and (c) Small Impact class (less than 10% variance). Consequently, only Large Impact class is used for the simulation study. Its high percentage of variance is significant and sufficient to comprehend the big picture of the events that might occur. Later, the months and dates in this class are used as a reference to mine the relevant news online.

(ii) Information Retrieval

Next, we retrieved the information that correspond to the contributing factors in this level. Research [12] retrieved information from stock market and used text mining to mine the news and analyse the correlations. In this research, text mining is also applied with Google News to retrieve the information. Google News is capable on rendering the results in specific timelines or date and this is very helpful. Nevertheless, proper selection of keywords for the news mining is essential to return definite and relevant news. As we are trying to understand the market and to retrieve relevant information regarding to it, 'Crude oil: price' is the keyword used to mine the definite news.

(iii) Feature and Rules Extraction

Afterwards, features related to crude oil market are extracted and analysed from the monthly news retrieved. The process of extracting the features from the news was done manually. Nonetheless, these extracted features helped us to understand the market. As mentioned, news is very useful information as it offers knowledge that contains rules. From the feature extracted, rules are collected and stored in a database to be used in the next model.

From these three processes, the key factors to crude oil market are discovered, verified and presented in Table 1. These key factors are then used as the input variables in subsection (*B).*

B. Artificial Neural Networks- Quantitative (ANN-Q) Model

ANN has gained much attention for its computational intelligence approach and its capability to make prediction. It is popular for capable on modelling the nonlinearity, which results to a class of general function approximators [13]. The development of this ANN-Q model is based on a process development suggested by [14] and presented as at Fig. 2. There are three steps of development for this model; (i) objective determination, (ii) data pre-processing and (iii) ANN modelling.

(i) Objective Determination

The objectives are determined to focus on developing a suitable and an accurate prediction tool for the crude oil market and predict its price for every barrel of crude oil in US Dollar (USD). USD is used as a standard price in NYMEX where WTI is traded. Moreover at this phase, we determined and analysed the selection of inputs and output to be used in the prediction so to find the fittest and right observations for the prediction. Nonetheless, it is important to ensure ANN capable of learning the connections between inputs effectively so to successfully achieve the final objective of the model. The networks' performance will be demoted if the selection of input and output of variables are not carefully selected. Simultaneously, the selection of input and output is a process of development and errors in mapping the correct input and output will make the prediction tool be less reliable [15].

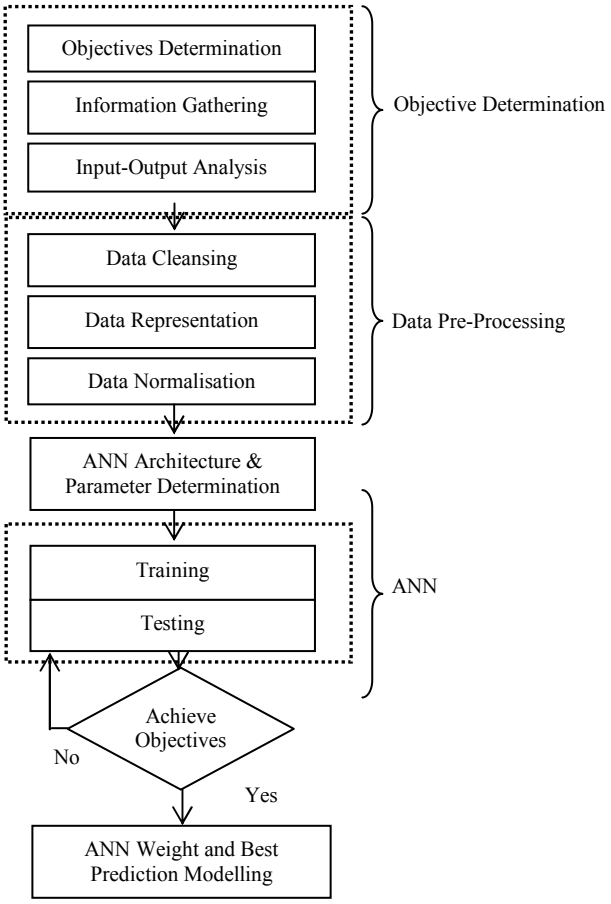

Fig 2. Artificial Neural Networks (ANN) Development Process

Based on information retrieved from HC model in (*A*), data collected and used for this model are 25 years of monthly quantitative data ranging from January, 1984 to February, 2009 sourced from Energy Information Administration (EIA), International Energy Agency (IEA), Economagic, World Energy, Population Reference Buraeu, World Bank and Source OECD. These data will be processed in the next phase to verify its fitting in the prediction model.

(ii) Data Pre-processing

All selected data in this phase are given extensive deliberation to ensure the cleanliness of data from noises and errors before further transformation into data representation stage. Furthermore, [16] suggests to collect data from reliable and premiere source, determine the series of time in specific range of date and have full cleansing on all data. These are to ensure no missing or inappropriate value of data present so to achieve fitter sets of data for the prediction. In this research, data used for training are time-series data and normalised data. Both of these types of data will be tested against each other to determine the best training model. The normalised data were represented with One-Step Returns function in equation (1),

$$R_n = \frac{x_n - x_{n-1}}{x_{n-1}} \quad (1)$$

where R_n is the monthly returns value for input number n with x and x_{n-1}denotes the input and past month observation respectively.

(iii) ANN Modelling

In this study, we used BPNN for training the input variables. Detail on BPNN can be referred to [17]. The networks can capture relatively complex environment as it contains multiple layers of interacting nonlinear neurons [11]. Nevertheless, the networks are composed by historical observations with assigned weight values for input and the future values as the output. The sample architecture for this network is as described in Fig. 3.

Input (x_1)	Input ($x..$)	Input (x_n)	Output (y_{+1})
2.80	2.51	2.87	2.60
2.60	2.24	2.34	2.80
.	.	.	.

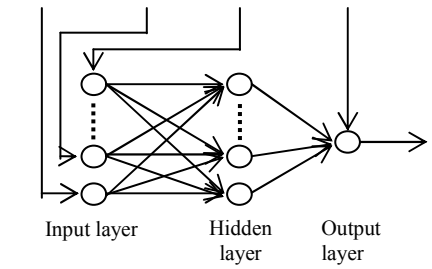

Fig. 3. The Sample Architecture of ANN for Prediction

Fig. 3 above shows the sample process in ANN where, the numbers of input and output variables selected are in parallel with the number of input and output layer in ANN. Selecting the right and appropriate numbers for hidden layers is important and its capability of reasoning will fail if extensive numbers of hidden layers are used in the model. In addition, the ideal technique for assigning the hidden layers can be obtained by referring to [18]. In this research, we used trial and error as our assigning method. We assigned 3, 4 and 5 layers to our networks as to make it simpler and more focus. Later, we test them individually to find the best fitting. This systematic process will ensure the accurateness of the final outcome and will validate the process we made in HC model.

Next, a simulation study is executed to validate the process done in (*A*) and (*B*).

V. Simulation Study

The ANN-Q model is composed by two modules; learning module and predicting module. In this section, we describe the data applied in this study together with discussion on both of the modules mentioned above. In this section, we also discuss on the evaluation metrics used to validate the performance and finally, evaluate the result from the simulation.

(i) Data Description

Input variables used in this simulation are the monthly historical data derived from quantitative key factors as

described in Table I with, monthly WTI price data used as the output. This 25 years duration of data is ranging from January, 1984 to February, 2009 compositing of 302 sets of data, 23 numbers of attributes and 6,946 observations. Additionally, these data are first derived from 1,100 online news mentioned in HC model. A value of 80% from total observation is used as training and the remaining as testing in learning module. Moreover, 20% of data are also used in predicting module as the input.

(ii) Learning Module

In learning module, data are represented in two types, One-step Returns function and time-series data itself. They are trained individually with the set of 3, 4 and 5 hidden layers, chosen by trial and error method. These sets of data are also trained with 8 different sets of input variables (refer to Table II) in the rationale of finding the smallest absolute error that will reflect the best correlation of inputs. This process is also in addition, to validate the key factors chosen in HC model. The inputs for this training are presented in Table II. Result from this learning module will be the input for predicting module in the next section.

(iii) Predicting Module

ANN-Q model learnt the rules from influential events occurred through the quantitative data provided in HC model.

TABLE II
THE TRAINING DATA SET FOR LEARNING MODULE

Training No.	Data Selection	Exceptional Input Data	No. of Input
1	All data	N/A	22
2	All except I^T	$I_{a1}, I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}$	16
3	All except E^T	$E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}$	16
4	All except P^T	P_{a1}, P_{a2}	20
5	All except ($I^T + E^T$)	($I_{a1}, I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}$) + ($E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}$)	10
6	All except ($I^T + P^T$)	($I_{a1}, I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}$) + ($P_{a1}, P_{a2}$)	14
7	All except ($E^T + P^T$)	($E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}$) + ($P_{a1}, P_{a2}$)	14
8	All except ($I^T + E^T + P^T$)	($I_{a1}, I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}$) + ($E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}$) + ($P_{a1}, P_{a2}$)	8

I^T, E^T and P^T refer to Table I

This learning process in particular helps the model to learn and reason every factor provided. Simultaneously, this process diagnoses rules contained in the quantitative data and delivers their prognosis. The objective of this module is to formulate a prediction result based on outputs produced from learning module. The example of prediction output for the first month; y_1 can be equated by (2).

$$y_1 = f(x_1 w_{11} + x_2 w_{21} + x_{..} w_{..1} + x_n w_{m1} - \theta_1) \quad (2)$$

In (2), $f(\ )$ represents the sigmoid function for the first month period of prediction where x denotes the input, w and θ denotes the weight and error (threshold) to the neuron respectively. The general equation for crude oil price can be represented as in (3) and (4). The denotation of these variables can be referred in Table I.

$$WTI_n = f(S^T + D^T + I^T + E^T + P^T - \theta) \quad (3)$$

$$WTI_n = f(S_{a1+a2+b1+b2+c1} + D_{a1+a2} + I_{a1+a2+b1+b2+c1+c2} + E_{a1+a2+a3+b1+c1+d1} + P_{a1+a2} - \theta) \quad (4)$$

The prognosis from this prediction is evaluated by three evaluation metrics to further validate the process and indicate the capability of ANN-Q to predict the crude oil price.

(iii) Performance Evaluation Metric

Performance evaluation metrics used in this research are Root Mean Squared Error (RMSE), Normalised Mean Squared Error (NMSE) and Directional Change Statistic (D*stat)*. The value of RMSE and NMSE are calculated by (5), (6) and (7). In addition, evaluation on the directional movement of predicted to actual data is evaluated by (8).

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \quad (5)$$

$$NMSE = \frac{MSE}{\frac{1}{N}\sum(\hat{y} - A)^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (7)$$

$$D_{stat} = \frac{1}{N}\sum_{t=1}^{N} a_t \quad (8)$$

Equation (5), (6) and (7) represents the calculation of NMSE and RMSE where N denotes the number of evaluation inputs, where y and $\hat{y}$ denotes the actual and predicted output respectively for time t. Concurrently in (8), a and t is expressed by,

$a_t = 1$ if $(y_{t+1} - y_t)(\hat{y}_{t+1} - y_t) \geq 0$, and $a_t = 0$ otherwise. D_{stat} is used to evaluate the predicted price movement. This performance metric is a good indicator and useful for practitioners to hedge their trading risk on crude oil market. The empirical result based on this simulation is discussed in the next section.

VI. EMPIRICAL RESULT

In this section, the empirical result from the simulation is presented and discussed. To begin, time series and normalised data are trained, tested and compared. The best result with the smallest absolute error value from this learning module will be the input data for prediction. From the simulation, we discovered the best learning data were derived from normalised data simulated with 5 hidden layers. This simulation shares promising 2.2690 of RMSE value, 0.00896 for its NMSE and finally, 93.33% for its D*stat* value. The prediction result for March, 2004 until February, 2009 is presented in Fig.4. This figure extensively shows narrow span between the actual and predicted price, expressing the accurateness of ANN-Q prediction model.

This accurateness not only implies to the trend but also to its discrete price. Therefore, it proves and validates the selection of variables chosen for the training. In addition, a parallel and positive movement existed between the actual price and the predicted price presented in Fig. 4 also validates the effectiveness of key factors selected in HC model. Furthermore, the figure shows that strong correlations existed between the factors.

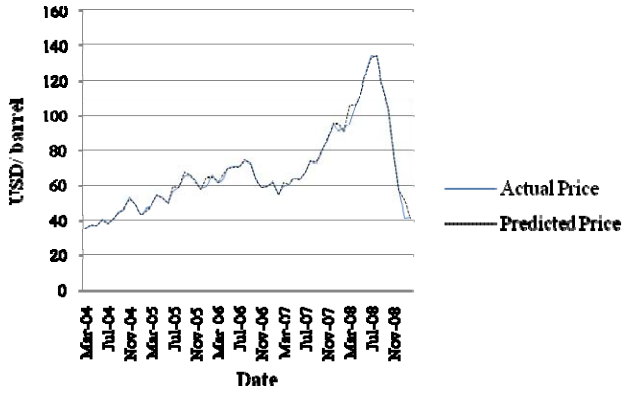

Fig. 4. Crude Oil Price Prediction Result Presented in One-step Returns Function Data

To further evaluate the performance, this study is assessed with two others recent hybrid prediction models by using RMSE and D*stat*. The two models compared are TEI@I Nonlinear Integration Forecasting model [10] and EMD-FNN-ALNN model [12]. Both of these models also used BPNN as their training tool. The comparison of these evaluation metrics are as at Table III.

TABLE III
CRUDE OIL PRICE PREDICTION PERFORMANCE EVALUATION

Methodology	RMSE	D*stat* (%)
TEI@I Nonlinear Integration model	1.0549	95.83
EMD-FNN-ALNN model	0.2730	86.99
ANN-Q model	2.2690	93.33

Based on Table III, the RMSE value for EMD-FNN-ALNN performed the first with 0.2730 followed by TEI@I Nonlinear Intergration and ANN-Q ranked the third place. The reason ANN-Q having the lowest RMSE in the simulation study is because of the normalised data used for the training. ANN-Q would perform better than the other two models if we are to compare its NMSE value. Based on that matter, it will also validate the importance of data pre-processing mentioned in [16] which was made absent in the other two models. According to RMSE value owned by EMD-FNN-ALNN, the decomposition and ensemble (or 'divide' and 'conquer') techniques used in [11], does support its claim as an alternative prediction tool for crude oil market.

Nonetheless, in terms of D*stat* comparison, *TEI@*I Nonlinear Integration has outperformed the other two models with 95.83% directional accuracy, followed by ANN-Q with 93.33% and lastly, EMD-FNN-ALNN with 86.99%. Even though ANN-Q performed the last for its RMSE value, the low RMSE does not necessarily mean that there is a high hit rate in predicting the crude oil price movement [19]. Hence, D*stat* is more practical in reflecting the fluctuate trend of the WTI price. However, ANN-Q is still a promising prediction tool for crude oil price with only 2.5% lesser than D*stat* winner; TEI@I Nonlinear Integration model. Albeit being a single model prediction, ANN-Q proved to be competitive and comparable to other prediction tools. This shows a positive opportunity for improvement in the near future.

VII. CONCLUSION & FUTURE WORK

In this study, machine learning and computational intelligence approach through HC and ANN-Q model is applied to predict the monthly WTI crude oil price for every barrel in USD. The result obtained from simulation study validates the effectiveness of data selection process by HC model. HC model successfully extracts a comprehensive list of key factors that cause the crude oil price market to volatile. The effectiveness and accurateness of data selection also helps to extensively deliberate the input variables combination for ANN-Q model. Data represented in One-step Returns function had successfully proved to cleanse and uniform the data from errors and noises hence, the crisp prediction result. This research is now in its extension level to comprehend this quantitative part of prediction with the qualitative part mentioned in part IV. This work in progress is expected to trigger and show some interesting information and trend for this crude oil price and together will result to a better prospect for crude oil price prediction in the future.

REFERENCES

[1] Y. Nelson, S. Stoner, G. Gemis, &, H. Nix, "Results od Delphi VIII Survey of Oil Price Forecasts," *Energy Report, California Energy Commisssion*, 1994.

[2] E. Panas, V. Ninni, "Are Oil Markets Chaotic? A Nonlinear Dynamic Analysis," *Energy Economics,* vol. 22, pp. 549-568, 2000.

[3] B. Abramson, A. Finizza, "Using belief networks to forecast oil prices," *International Journal of Forecasting* vol. 7, 1991, pp. 299-315.

[4] B. Abramson, A. Finizza, "Probabilistic forecasts from probabilistic models: A case study in the oil market," *International Journal of Forecasting* vol. 11, 1995, pp. pp. 63-72.

[5] C. Morana, "A Semiparametric Approach to Short-term Oil Price Forecasting," *Energy Economics*, vol. 23, 2001, pp. 325-338.

[6] G. Barone-Adesi, F. Bourgoin, & K. Giannopoulos, "Don't Look Back," *Book Don't Look Back*, Series Don't Look Back,1998, pp. 100-103.

[7] M. Ye, J. Zyren, & J. Shore, "A Monthly Crude Oil Price Forecasting Model Using Relative Inventories," *International Journal of Forecasting*, vol. 21, 2005, pp. 491-501.

[8] W. Xie, L. Yu, S. Xu, S. Wang, *A New Method for Crude Oil Price Forecasting Based on Support Vector Machines* Springer Berlin / Heidelberg, 2006.

[9] S. Wang, L. Yu, & K.K. Lai, "Crude Oil Price Forecasting With TEI@I Methodology," *Journal of Systems Science and Complexity*, vol. 18, no. 2, 2005, pp. 145-166.

[10] L. Yu, S. Wang & K.K. Lai, "A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting," *International Journal of Knowledge and Systems Sciences*, vol. 2, no. 1, 2005.

[11] L. Yu, S. Wang & K.K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm " *Energy Economics*, vol. 30, no. 5, 2008, pp. 2623-2635.
[12] P. Kroha, R. Baeza-Yates, &, B. Krellner, "Text Mining of Business News for Forecasting," *Proc. Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)*, 2006, pp. 1-5.
[13] Z. Tang, P.A. Fishwick, "Feedforward Neural Networks as Models for Time Series Forecasting," *ORSA Journal on Computing, 1993*, vol. 5, no. 4, 1993, pp. pp. 374–385.
[14] V. Rao, H. Rao, "C++ Neural Networks Application in Financial Asset Management," *Neural Computing and Application Journal* vol. 2, 1995, pp. 13-39.
[15] E.M. Azoff, *Neural Network Time Series Forecasting of Financial Markets*, John Wiley & Sons, 1994.
[16] M.F. Nasrudin, "*Pembangunan Model dan Aplikasi Ramalan Pasaran Saham BSKL Menggunakan Rangkaian Neural Perambat Balik*," Msc. of Computer Science, Faculty of Technology and Science Information, Universiti Kebangsaan Malaysia, Bangi, Selangor, 2001.
[17] M. Negnevitsky, *Book Artificial Intelligence: A Guide to Intelligent System*, Pearson Education Ltd., 2005, p. 415.
[18] C.N.W.W. Tan & G.E. Wittig, "A study of the parameters of a backpropagation stock priceprediction model," *Proc. International Two-Stream Conference Proceedings on Artificial Neural Networks and Expert Systems*, 1993, pp. 288-291.
[19] S. Wang, L. Yu & K.K. Lai, "An EMD-Based Neural Network Ensemble Learning Model for World Crude Oil Spot Price Forecasting," *Soft Computing Applications in Business*, no. 230, 2008, pp. 261-271.