CHAPTER 3

RESEARCH METHODOLOGY

Chaigusin (2011) mentioned that stock markets have different characteristics, depending on the economies they are related to, and, varying from time to time, a number of non-trivial tasks have to be dealt with when developing Neural Networks for predicting exchanges. It is not easy task to design artificial neural network model for a particular forecasting problem or a stock market index movement. Therefore, Modelling issues must be considered carefully because it affects the performance of an ANN. One critical factor is to determine the appropriate architecture, the number of optimal hidden layers as well as the number of hidden nodes for each layer. Other network design decisions include the selection of activation functions of the hidden and output nodes, the training algorithm, and performance measures. The design stage involves in this study to determine the input (independent) and output (dependent) layers through the hidden layers in the case where the output layer is known to forecast future values. Output of the network was two patterns *0* or *1* of stock price direction. The output layer of the network consisted of only one neuron that represents the direction of movement. The number of neurons in the hidden layer was determined empirically. The determination of the formulation between input and output layers is called learning and through the learning process, model recognises the patters in the data and produces estimations.

From the literature, multi-layer feed-forward ANN with back-propagation is the most commonly used architecture in this area. So, we use the three-layered feed-forward architecture (see Fig. 2). The entire data set covers the period from 03/01/2005 to 30/12/2010 for network training, while data from 03/01/2005 to 28/05/2014 is to test the predictive ability of the network.

There are some steps as follow:

12 indicators has to be calculated in excel and then the results will be loaded to the network for training and testing,

The data will be loaded to the network and then Normalization will take place ranging between -1, 1 so that the network will able to learn faster, training period will be in yearly because of avoiding too much of time consuming.

Training process will take place within time frame (20 minutes), if the process cannot reach the goal, and then changing its learning rate and momentum constant will be needed.

Looking for the best parameter combination that enhance the best output and save as “net” for testing step(forecasting)

The testing process can be conducted in the new set of data to see how best the performance of the model

The basic methodologies applied in this research are based on previous researches such as (Kim 2003, Mahmood Moein Aldin et al. 2012, Najeb

Masoud 2014,...) The performance evaluation of the model can be described below:

3.1 Statistical Performance Evaluation of the Model

In order to estimate the forecasting statistical performance of some methods or to compare several methods we should define error functions. Many previous research works had applied some of the following forecast accuracy measures: Mean Error (ME), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Standard Deviation of Errors (SDE), Mean Percent Error (MPE) and Mean Absolute Per cent Error (MAPE), etc. In our study we use four performance criteria namely mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and goodness of fit R ^{2}. The back-propagation learning algorithm was used to train the three-layered feed-forward ANN structure in this study were the most used error functions is as following:

The mean absolute error is an average of the absolute errors E = (P_i - $\hat{p}_i$), where P_i and $\hat{p}_i$ are the actual (or observed) value and predicted value, respectively. Lesser values of these measures show more correctly predicted outputs. This follows a long-standing tradition of using the “ex-post facto” perspective in examining forecast error, where the error of a forecast is evaluated relative to what was subsequently observed, typically a census based benchmark (Poon 2005). The most commonly used scale-dependent summary measures of forecast accuracy are based on the distributions of absolute errors (|E|) or squared errors (E^2) observations (n) is the sample volume. The mean absolute error is given by:

$$\text{Mean Absolute Error (MAE)} = \left(\sum_{i=1}^{n}\left(\left|E\right|\right)/n\right) \qquad (i = 1, 2,\ldots n) \qquad (6)$$

The *MAE* is often abbreviated as the *MAD* (“D” for “deviation”). Both *MSE* and *RMSE* are integral components in statistical models (e.g., regression). As such, they are natural measures to use in many forecast error evaluations that use regression-based and statistical. The square root of the mean squared error as follows:

$$\text{Mean Square Error (MSE)} = \left(\sum_{i=1}^{n}\left(\left|E^2\right|\right)/n\right) \qquad (i = 1, 2,\ldots n)$$

$$\text{Root Mean Square Error (RMSE)} = Sqrt\left(\sum_{i=1}^{n}\left(\left|E^2\right|\right)/n\right) \quad (i = 1, 2,\ldots n) \quad (7)$$

If the above RMSE is very less significant, the prediction accuracy of the ANN model is very close to 100%. Since percentage errors are not scale-independent, they are used to compare forecast performance across different data sets of the area using absolute percentage error given by APE = (P_i - $\bar{P}_i$)*100. Like the scale dependent measures, a positive value of APE is derived by taking its absolute value (| *APE* |) observations (n). This measure includes:

$$\text{MAPE} \quad = \quad \left(\sum_{i=1}^{n}\left(\left|APE\right|\right)/n\right) \qquad (i = 1, 2,\ldots n) \qquad (8)$$

The use of absolute values or squared values prevents negative and positive errors from offsetting each other. All these features and more make MATLAB an indispensable tool for use in this work.

$$\text{Goodness of Fit } (R^2) = \left(\sum_{i=1}^{n}(E^2)/(e^2)\right) \qquad (i= 1, 2,\ldots n) \qquad (9)$$

where $e_i = p_i - \overline{p}_i$, is the forecast error values. p_i, the actual values and $\overline{p}_i$, denote the predicted values. The more R^2 correlation coefficient gets closer to one, the more the two data sets are correlated perfectly. As the aim of all of the prediction system models proposed in this study is to predict the direction of the stock price index forecasting, the correlation between the outputs do not directly reflect the overall performance of the network.

3.2 Financial Performance Evaluation of the Model

In order to evaluate the financial performance of the model, the correct predicted positions by the model have been compared. Prediction performance is evaluated used in the formula to calculate the prediction accuracy (Kim 2003) and is as follows:

$$Prediction\ (P) = \frac{1}{n}\sum_{i=1}^{n} R_i \qquad (i = 1,2,...n) \qquad (10)$$

Where $\mathbf{R}_i$ the prediction result is for the $\mathbf{i}^{th}$ trading day is defined by:

$$R_i = \begin{cases} 1 & If\ \ PO_i = AO_i \\ 0 & Otherwise \end{cases}$$

PO_i is the predicted output from the model for the i^{th} trading day, and AO_i is the actual output for the i^{th} trading day, **n** the total predicted outputs. The error level was determined 5% and it means that those outputs with the error level less than the defined value are considered as correctly predicted values.

3.3 Research Data

The research data used in this study is the direction of change in the daily Jakarta composite stock price index (JKSE). This is composed of closing price, the high price and the low price of total price index. The grand total number of sample is 2,298 trading days, from January 3, 2005 to May 28, 2014. It is divided into two sub-periods. First sub-periods of January 3, 2005 to December 30, 2010 is in network training periods, its values are obtained with different combinations of parameters for testing the models. The second sub-period of January 3, 2005 to May 28, 2014 is in sample period for testing prediction rate. The whole data in the statistical population were employed in the analysis and this leads to non-selection of a specified sampling method. The number of sample with increasing direction is 1,303 while the number of sample with decreasing direction is 995. That is, 57% of the all sample have an increasing direction and 43% of the all sample have a decreasing direction. The research data used in this study is the direction of daily closing price movement in the JKSE. The number of sample for each year is shown in Table 1.

Table 1. The number of sample in the entire data set

Description	Year										Total
	05	06	07	08	09	10	11	12	13	14 May	
Increase	136	144	151	123	141	140	137	135	131	62	1,300
(%)	56%	59%	60%	51%	58%	57%	55%	55%	55%	63%	57%
Decrease	107	101	109	120	102	105	110	109	109	36	998
(%)	44%	41%	40%	49%	42%	43%	45%	45%	45%	37%	43%
Total	243	245	250	243	243	245	247	244	240	98	2,298

Source: author calculation, 2014

3.4 Data preparation

Some data own a high amount in comparison with others and this might lead to the excessive effect on prediction process which is a source of errors and reduction of prediction ability of neural networks. That’s why the original data should be normalized in a range of [l, h]. with regards to Mahmood Moein Aldin et al. (2012) normalizing data is done as follows:

$$u = \frac{(x_i - x_{i,min})}{(x_{i,max} - x_{i,min})}(h_i - l_i) + l_i \qquad (i = 1,2,...n) \quad (11)$$

Where:

u = the normalized data

x_i = the original data

$x_{i,min}$ = the minimum value of the original

$x_{i,max}$ = the maximum value of the original data

h_i = upper bound of the normalising interval and

l_i = lower bound of the normalising interval

Max-min normalization plans a value **u** of $\mathbf{x_i}$ in the range ($\mathbf{h}_i$ – $\mathbf{l}_i$) i.e. (-1.0; 1.0), in this case. As a value greater than 0 represents a buy signal while a value less than 0 represents a sell signal. (i = 1,2,3,...,n) the number of observations.

3.5 Variable Calculation

Closing price, the high and low price index are converted into technical indicators. Technical indicators are used as input variables in the construction of prediction models to predict the position of stock price movements. In this research, 12 technical indicators has to be calculated in Excel and then the network (Program matlab) will read the results from excel spreadsheet. Training or learning data will be year on year, because if we combine data of many years to train at one time it means the learning process is very long and sometimes may not reach the goal.

The research applied indicators are selected based on indicator selection of different groups and also along with the previous studies Kim (2003), Kumar & Thenmozhi (2006), Kara et al. (2011), Mahmood Moein Aldin et al. (2012), A. Victor Devadoss (2013)… Table 2 demonstrates the titles of twelve technical indicators and their calculation method separately.

Table 2. Selected technical indicators and their formulas

No	Name of indicators	Formulas	Description
1	A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$	where C_t is the closing price at time t, L_t the low price at time t, H_t the high price at time t (J. Chang et al. 1996)
2	CCI Commodity Channel index	$\frac{M_t - SM_t}{0.015D_t}$	Where $M_t = (H_t + C_t + L_t)/3$, $SM_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$ $D_t = \frac{\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert}{n}$ (S.B. Achelis, 1995 & J. Chang et al. 1996)
3	Larry William's (R%)	$\frac{H_n - C_t}{H_n - L_n} * 100$	(S.B. Achelis, 1995)
4	MACD (moving average convergence divergence)	MACD(n)$_{t-1}$+2/n+1* (DIFF$_t$ - MACD(n)$_{t-1}$)	DIFF: EMA(12)$_t$ - EMA(26)$_t$, EMA is exponential moving average, EMA(k)$_t$: EMA(k)$_{t-1}$ + α ×(C$_t$ - EMA(k)$_{t-1}$), α smoothing factor: 2/(1 + k), k is time period of k day exponential moving average (Gerald, 2005)
5	Momentum	$C_t - C_{t-n}$	where C_t is the closing price at time t, n the price day (J. Chang et al. 1996)
6	ROC Price-rate-of change	$\frac{C_t}{C_t - n} * 100$	(J.J. Murphy, 1986)
7	RSI (Relative strength index)	$100 - \frac{100}{1 + \left(\sum_{i=0}^{n-1} Up_{t-1}/n\right)/\left(\sum_{i=0}^{n-1} Dw_{t-}\right.}$	where Up_t means upward-price change and Dw_t means downward price-change at time t. (S.B. Achelis, 1995)

No	Name of indicators	Formulas	Description
8	Simple MA	$\frac{C_t + C_{t-1} + \cdots C_{t-n}}{n}$	It shows the average value of a security's price over a period of time. If the value of a security's price over a period of time. If the price moves above its MA, a buy signal is generated. If the price moves below its MA a sell signal is generated. (Mahmood Moein Aldin et al. 2012 & Najeb Masoud, 2014)
9	Stochastic *(K %)*	$\frac{C_t + LL_{t-n}}{HH_{t-n} - LL_{t-n}} * 100$	where LL$_t$ and HH$_t$, mean lowest low and highest high in the last t days, respectively. (S.B. Achelis, 1995)
10	Stochastic *(D%)*	$\frac{(\sum_{i=0}^{n-1} K_{t-i} \%)}{n}$	(S.B. Achelis, 1995)
11	Stochastic slow (D%)	$\frac{(\sum_{i=0}^{n-1} D_{t-i} \%)}{n}$	(E. GiEord, 1995)
12	WMA	$\frac{(n) * C_t + (n-1) * C_{t-1} + \cdots + C_n}{(n + (n-1) + \cdots + 1)}$	Mahmood Moein Aldin et al. (2012), Najeb Masoud (2014)

Notes: In this study the original data were normalized in a range of [-1,1].

Table 3. Defined Variables

Code	Definitions
A/D Oscillator	Accumulation/distribution oscillator. It is a momentum indicator that associates changes in price
CCI Commodity Channel index	It measures the variation of a security’s price from its statistical mean
Larry William’s (R%)	It is a momentum indicator that measures overbought/ oversold levels
MACD (moving average convergence divergence)	Moving average convergence divergence
Momentum	It measures the amount that a security’s price has changed over a given time span
ROC Price-rate-of change	It displays the difference between the current price and the price n days ago
RSI	Relative strength index. It is a price following an oscillator that ranges from 0 to 100. A method for analysing RSI is to look for divergence in which the security is making a new high.
Simple MA	Simple 10-day moving average
Stochastic *(K %)*	It compares where a security’s price closed relative to its price range over a given time period
Stochastic *(D%)*	Moving average of %K
Stochastic slow (D%)	Moving average of %D.
WMA	Weighted 10-day moving average

Source: Kim K. (2003), Kara et al. (2011), Mahmood Moein Aldin et al. (2012), Najeb Masoud (2014)