Neural Network with Genetic Algorithm Prediction Model of Energy Consumption for Billing Integrity in Gas Pipeline

By

AIDIL FAZLINA BINTI HASBULLAH

FINAL PROJECT REPORT

Submitted to the Department of Electrical & Electronic Engineering
in Partial Fulfillment of the Requirements
for the Degree
Bachelor of Engineering (Hons)
(Electrical & Electronic Engineering)

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Neural Network with Genetic Algorithm Prediction Model of Energy Consumption for Billing Integrity in Gas Pipeline

by

Aidil Fazlina Binti Hasbullah

A project dissertation submitted to the
Department of Electrical & Electronic Engineering
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
Bachelor of Engineering (Hons)
(Electrical & Electronic Engineering)

Approved:

_____

Ir. Dr. Idris Bin Ismail
Project Supervisor

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

June 2012

i

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

_____

Aidil Fazlina Binti Hasbullah

ABSTRACT

Along the development of oil and gas industry, missing data is one of the contributors that restrains in analyzing and processing data task in database. By monitoring and maintaining using metering system, the reliability and billing integrity can be ensured and trustworthy can be developed between distributors and customers. In this context, PETRONAS Gas Berhad (PGB) as a gas distributor and an existing system in Nur Metering Station, Kulim is held responsible to evaluate the energy consumption from the sales gas produced. The system is standalone that consists of measuring equipment including pressure transmitter and temperature transmitter, turbine meter, gas chromatography and flow computer but does not have any reference system to verify its integrity. Customers are being charge according to the amount of energy consumption calculated and any error in calculation will cause loss of profit to the company and affect PETRONAS's business credibility. In the future, it is such a vital to have an ideal analysis in order to maintain the sustainability. In this paper, several techniques will be discuss and selected including neural network prediction model, least square vector regression and combination of either two methods mentioned before with genetic algorithm as preferable technique to indicate the missing data. The model that has been selected based on its evaluation will predict the missing data and compare it with the results of the existing metering system to ensure the reliability and accuracy of the system. The billing integrity between oil and gas company especially PETRONAS and the customers could be maintained and in the future if the project is expanded, it will have the potential of saving of millions of dollars to Malaysian oil and gas companies.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude towards Allah the Almighty for blessing me with vigour and great health throughout the period in order to complete this project. The utmost thankfulness goes to my supervisor, Ir. Dr Idris Bin Ismail for his continuous guidance and support throughout my project. Another appreciation I would like to give towards Mr. Yisa Ademola Adeeyo, who has been a great teacher, and helping me getting through many ups and downs during this project. Besides that, I would like to thank Universiti Teknologi PETRONAS (UTP), especially the Electrical and Electronics Engineering Department whereby students are given the opportunity to broad up their knowledge and experience the real challenges of managing research and project before entering such a crucial world of working. It is an opportunity for us, students to use the tools and techniques for problem-solving. Furthermore, the learning process is gained through hands-on experience. Last but not least, thank you to all my fellow friends and family for their love and support, I have successfully completed my Final Year Project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

TOD	Transmission Operation Division
PGB	PETRONAS Gas Berhad
AGA	American Gas Association
MCAR	Completely at Random
MAR	Missing at Random
STLF	Short Term Load Forecasting
MTLF	Medium Term Load Forecasting
LTLF	Long Term Load Forecasting
ANN	Artificial Neural Network
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Squared Error
LS-SVR	Least Squares Support Vector Regression
SVR	Support Vector Regression
RSVMG	Recurrent Support Vector Machines with Genetic Algorithms
GUI	Graphical User Interface

LIST NOMENCLATURE

E_p	Measure of error on pattern
t_{pj}	Target output for j-th order
o_{pj}	Actual output for j-th order
ξ	Slack variables
ξ^*	Slack variables
C	Regularization parameter
α_i	Langrange multiplier
$\vec{X}$	Input vector
$\vec{Y}$	Output vector
$\vec{W}$	Vector of weights
e	Error variable
ε	Mean Absolute Percentage Error

CHAPTER 1
INTRODUCTION

1.1   Background Study

To be compatible with such a rapid competition in the oil and gas industry, missing and incomplete data will creates various tribulations in analyzing and processing data. Natural gas load forecasting seems to be the solution to the efficient and profitable for pipeline network operation. Pipelines are used for the transportation of natural gas to the consumers. To achieve the standards of prerequisite for the quality of automatic control in the process industries, it leads to the opportunity of system design with intelligent function that can measure and monitor the outgoing gas flow. The system must be ensured reliable and dependable at all cost. The main concern is the products are to establish an accurate data that have been missed along the process.

An existing system is constructed in Transmission Operation Division (TOD), PETRONAS Gas Berhad (PGB) in order to calculate the energy consumption from the sales gas produced. The system is standalone that consists of measuring equipment including pressure transmitter and temperature transmitter, turbine meter, gas chromatography and flow computer but does not have any reference system to verify its integrity. But for reference of missing data is yet unavailable for the system hence, PGB could not detect specifically whenever issues occur.

Figure 1-1: Metering area in Metering Station
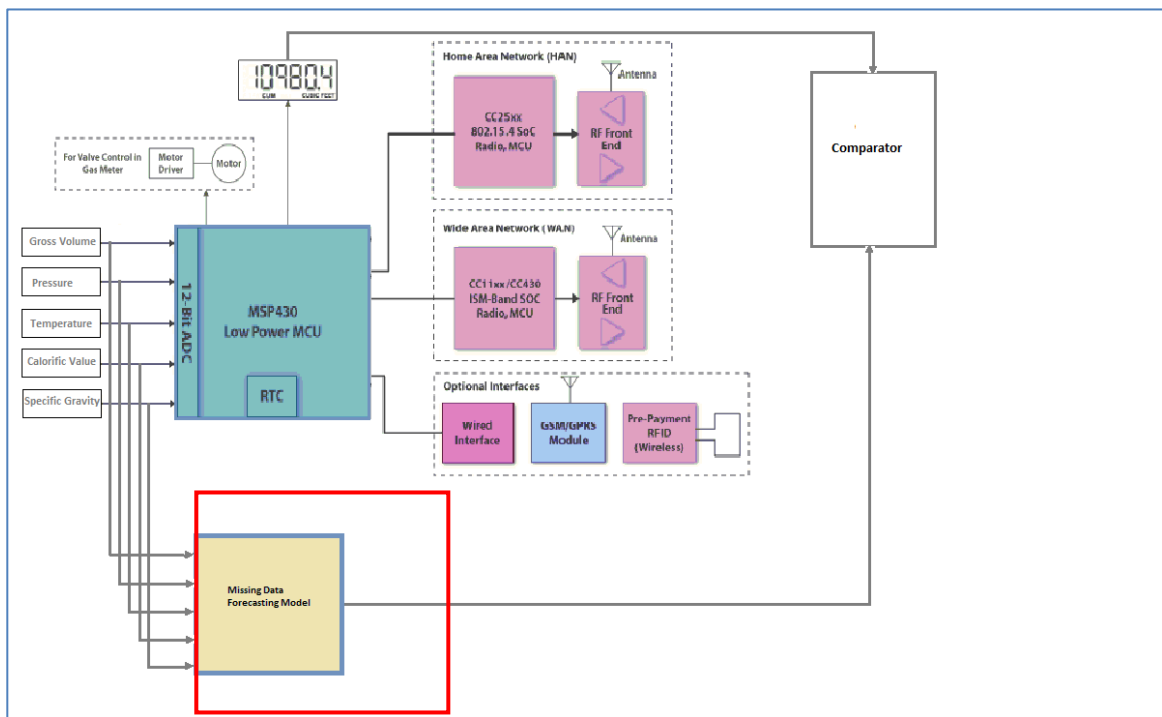
Figure 1-2: Metering Block Diagram Prediction Model Suggested

In order to do so, various method of forecasting will be discuss later to select the best method in terms of reliability, accuracy and sustainability. The methods are including artificial neural network integration, least-squares support vector regression and combination of both methods with genetic algorithm. In next chapter, discussion in terms of advantages, process procedures and implication of all methods will be

elaborated with various examples. The selected prediction model then will be utilized to predict the energy consumption of the sales gas in gas transmission metering in oil and gas sector especially for PETRONAS for ensuring a reliable and accurate metering system for billing integrity.

1.2   Problem Statement

According to J. Paris and R. Ibrahim, The existing metering system at TOD, PGB measuring and monitoring the energy consumption of the sales gas produced using the American Gas Association (AGA) equation [10]. The inputs for the calculation are obtained from all of the measuring equipments. The inputs are pressure, temperature, gross volume, calorific value and specific gravity. However, in actual condition, temperature and pressure varies. In this case, there will be a slight difference between the calculated energy consumption and the actual energy consumption during gas transmission between PGB and the customers. This may lead to error existence through the process and such a feasible model should be developed in order to maintain the efficiency of the production unit. The continuation of J. Paris and R. Ibrahim project will be done in this research in order to enhance the development of the project reliability.

1.3   Project Objectives

The objectives of this project are listed as below:

- To enhance the development of an existing prediction model of energy consumption that is based on a few process variables such as temperature, pressure, flow rate and characteristics of natural gas.
- To verify and compare among prediction model of energy consumption that come out with the most reliable result as a verification tool of the outgoing gas flow.
- To develop a prediction model that will produce accurate readings that is less than one percent error of energy prediction.

3

1.4   Scope of Study

The selected prediction model of energy consumption will be utilized as a verification tool for billing integrity of the existing metering system. The inputs of the model are based on the parameters from the actual plant which are pressure (P), temperature (T), gross volume (V_{gs}), calorific value (CV) and specific gravity (sg) while the output for the metering system is energy consumption (E). The model should be able to forecast the energy consumption based on the inputs and the predicted energy will be compared with the actual energy.

The selected approach to construct the model is using MATLAB software. The software has some functions that are found feasible for the project at this phase. The selected model will learn the association between input parameters and controlled and uncontrolled variables by studying the previous recorded data. The output will be predicted by the model based on the trained data earlier for other input.

1.5   Relevance of Project

The project is proposed to ensure the reliability and accuracy of the existing metering system in (PGB) in the hope to maintain the billing integrity between PGB and the customers. A successful model will reduce dispute between PETRONAS and the customers and at the same time PETRONAS will get to maintain its business credibility and loss of profit to the company could be prevented. The customers would also gain confidence and trust on business dealing with PETRONAS.

1.6   Feasibility of Project

The project is based on the actual situation in TOD, PGB and a preliminary investigation has been done by an engineer from PGB. The idea of the investigation is to find the most suitable and promising method to be proposed because based on previous research done by J. Paris and R. Ibrahim, there are slightly flaw while using artificial neural network [10].

The project is scheduled to be completed within two semesters. The project can be completed within the scope and the timeline as long as the individual responsible for the project gives full commitment and strive to get the most optimum results and present a successful project at the end.

CHAPTER 2

LITERATURE REVIEW

2.1   Prediction Models of Missing Data for Natural Gas Consumption

In such a large industries of oil and gas, every angle of process should be as accurate as possible. Missing data will create various issues in analyzing and processing data in databases [1]. According to P. D. Allison[2], the types of missing data mechanism can be categorized into three which are Missing Completely at Random (MCAR), Missing at Random (MAR) and non-ignorable. MCAR is defined whenever the probability of missing data for variable X is independent to the value X itself or to any other variable in the data set occurred. On the other hand, MAR happens if the chance of missing data on a particular variable X independent on X but depends on other variables, and the non-ignorable case achieved if even the other variables is under control by the user, the probability of missing data X is still related to the value of X itself in the analysis. The methods discussed in this paper should be applicable to all situations have been stated before.

For this purpose, the methods for predicting natural gas consumption can be divided into different categories, such as intelligent and non-intelligent. In this paper, review on each category and comparison among the categories will be done in order to select the best method of forecasting model. Under the intelligent forecasting model which is the focal point on the development of the predicting model in this project, further discussion upon artificial neural network, least square vector regression and genetic algorithm. To select the best model, consideration upon period of sample data also will be made which are Short Term Load Forecasting (STLF), Medium Term Load Forecasting (MTLF) and For Long Term Load Forecasting (LTLF) [3].

6

There are several techniques for load forecasting data captures duration in order to ensure the reliability of the accurate results and regarding to the operation of the process system. Load forecasting can be divided into three categories which are short term, medium term and long term [3]. STLF indicates the duration for prediction model to capture the data is usually from one hour to a week [4]. STLF also offer a great economic saving potential and secure operation of power system. On the other hand, MTLF means the load forecasting usually take a longer period of time which is from a week to a year [5]. Furthermore, MTLF usually influence several additional factors especially demographic and economic [4]. LTLF is usually more than a year [5] which is very useful for a longer duration of planning operations.

2.2   Volume and Energy Calculation

As stated before in the introduction, for this project, the gas volume is determined from metered volume monitoring, pressure and temperature measurements from the transmitter and gas composition data. Based on American Gas Association Report No.8 and ISO 6979, the energy consumption of sales gas is being determined as stated in the equation below describing the compressibility factor standard and analysis. This equation is based on hourly volume of gas (standard cubic meter) for each hour.

$$V_h = \sum_h \left(\frac{V_{g(meas)} \times P_{meas} \times T_{base} \times Z_{base}}{P_{base} \times T_{meas} \times Z_{meas}} \right)$$

(2.1)

Normally, the main task in the gas industry is to provide the energy measurement. The complexity of energy determination of gas is far beyond the measurement and billing of electrical energy and requires greater effort to achieve it. It is because fluid mechanics and calorific properties of natural gases have to be taken into account [27]. Below is the equation that being used for calculation of hourly energy flow (in GJ) for each hour,

7

$$E_h = \frac{(V_h \times CV)}{1000}$$

<div align="right">(2.2)</div>

$E_h =$ Energy consumption per hour (GJ)

$V_h =$ Volume of gas per hour (Sm^3)

$CV =$ Calorific value (mJ/Sm^3)

$V_{g(meas)} =$ Measured gas volume (m^3)

$P_{meas} =$ Measured pressure (kPa)

$P_{base} =$ Base pressure (kPa) (101.325 kPa)

$T_{meas} =$ Measured temperature (K)

$T_{base} =$ Base temperature (K) (288.15K)

$Z_{meas} =$ Compressibility factor at P_{meas} and T_{meas} calculated by reference to AGA Report No. 8

$Z_{base} =$ Compressibility factor at P_{base} and T_{base} calculated by reference to AGA Report No. 8

2.3   Artificial Neural Network

Intelligent prediction models require the system to have the ability to learn on a base of recall the previous data provided. The area of concern includes the process plant, environment of the system, controller and new aims and limit which is in our concern here is the missing data [7]. Artificial Neural Network (ANN) Method is one of the intelligent prediction models that usually applied to predict the short-term load for a large process system. By using ANN as prediction model, to control highly uncertain, non-linear and complex systems is one of the beneficial that can be achieved. By

allowing the unknown nonlinearities to be controlled in the system, the division of uncertain systems that adaptive control techniques can be applied in ANN application [8].

2.3.1 Artificial Neural Network Architecture

In an ANN, the main element of the process is also known as neurons. These neurons are uniform by means of information channels that being connected to each other called interconnections. Each neuron can have multiple inputs in order to determine only one output. The input from a neuron can be defined in several ways, either it could be from the external stimuli or we can say input inserted by the user and another way is from output of the other neurons where it shows the interconnection. The feedback connection is when clone of the single output that comes from a neuron also could be input to many other neurons in the network. There is connection strength, synapses, or weight associated with each connection [7]. When the neuron is excited because of the weighted sum of the inputs to the neuron exceeds a certain threshold, an output signal is produced. The backpropagation learning algorithm is the most frequently used method in training the networks instead of Hebbian and Comparative and will be focused in the discussion for this paper.
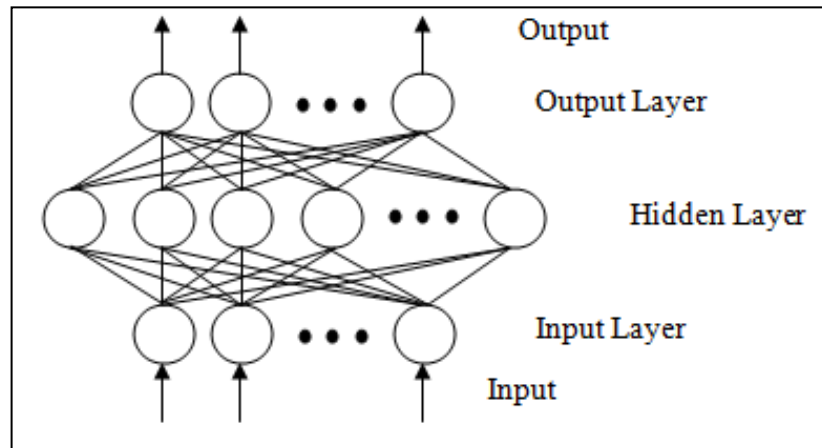
Figure 2-1: Schematic of Neural Network

From the generalization of Widrow-Holf error correction rule, the backpropagation learning is being derived. The Widrow-Hoff basically is an error signal formation technique to determine the difference between the actual outputs is and theoretical output. The synaptic strengths, or weights, were changed along in the proportion to the regression of error times the input signal, until which diminish the error in the direction of the gradient. But the existence of too many synapses can increase the value of error and complication occur not just the ones at the output units.
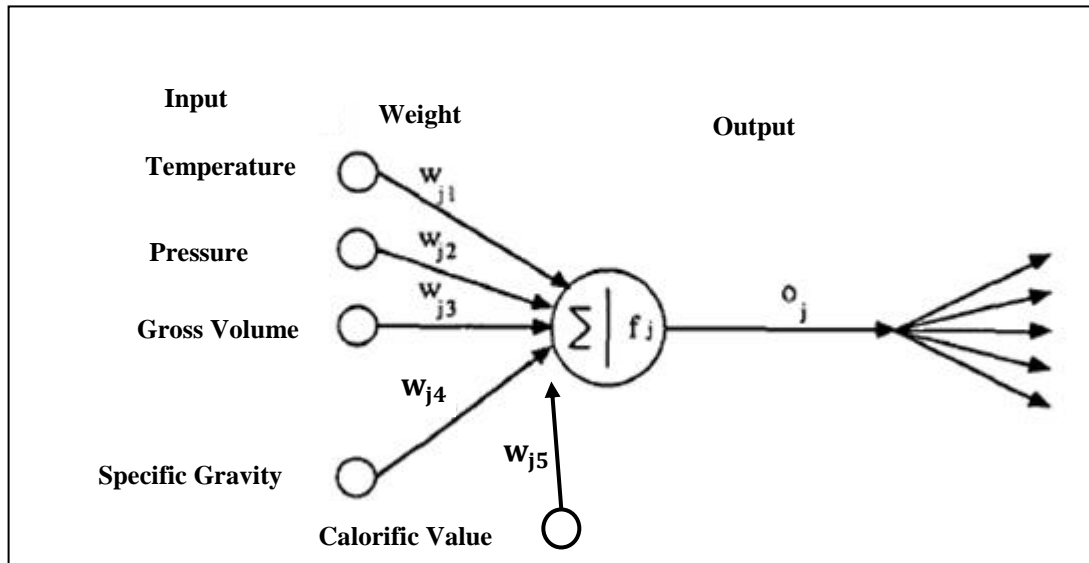
Figure 2-2: Mathematical Model of Neuron

"Generalized delta rule" is suggested by Rumelhart et al to gives a formulae for adjusting the weights automatically on internal units regarding to the error determined between predictions an actual at the output. As below, let [7];

$$E_p = \frac{1}{2}\sum_j (t_{pj} - o_{pj})^2,$$

(2.3)

10

Where E_p is the value of pattern error and $E = \sum_p E_p$ is the sum of pattern error with the respective t_{pj} is the target output for j-th component and o_{pj} is the actual output of j-th component. The error analysis is important for this forecasting technique hence Root Mean Square Error (RMSE), is the formulae for after-the-fact error analysis [8]:

$$RMSE = \sqrt{\frac{1}{j}\sum_{i=1}^{j}(t_{pj} - o_{pj})^z}$$

(2.4)

The procedure for the designed ANN model will be brief below in order to make it clear how exactly the proposed method will be conducted [8].
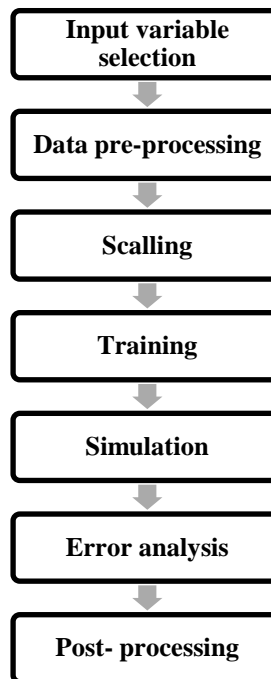
Figure 2-3: ANN-based Load Forecasting Procedure

First step is input variable selection which in this case is consisting of five inputs where the identification factors take place for making an acceptable forecasting. Then data will be re-processed in order to record the data and observation errors are inevitable.

11

Statistical method is used to identify and filter the bad and irregular data to be adjusted or discard. Later, in order to converge problems occurred maybe caused by direct use of input lead to different ranges for variables, scaling method will be used. When ANN is started to learn, system initializes weight and biases of each layer. Network will identify the appropriate transformation and mutation between past inputs and outputs from training cases. This level is when training level will take part. After training session being done, simulation upon forecasting output is being done using training ANN. When the characteristic is varying, error analysis is important for process of forecasting. Thus, RMSE is used to find after-the-fact error analysis. ANN output needs to be rescaled for generating preferred forecasted loads in post-processing.

2.3.2   Application Examples

By referring to some examples on application of ANN, there will be some review on the selected papers which have high citation. The selection papers will be based on the objective itself which is the beneficial of the implementation of ANN in load forecasting. Those papers are listed as follow:

2.3.2.1 A Short term streamflow forecasting using artificial neural networks [9]

The research described in this article investigates the functionality of ANN for short term forecasting of streamflow into Namakan Lake by C. M. Zealand et al. The area of focus is of approximately 20 000 km² stremflow using quarter monthly time intervals. From the research and simulation, during the calibration (training) phase there is a very close fit was attained then the ANNs was developed and outperformed a conservative model during verification (testing). The performance is being analyzed using root mean squared error (RMSE).

Result:
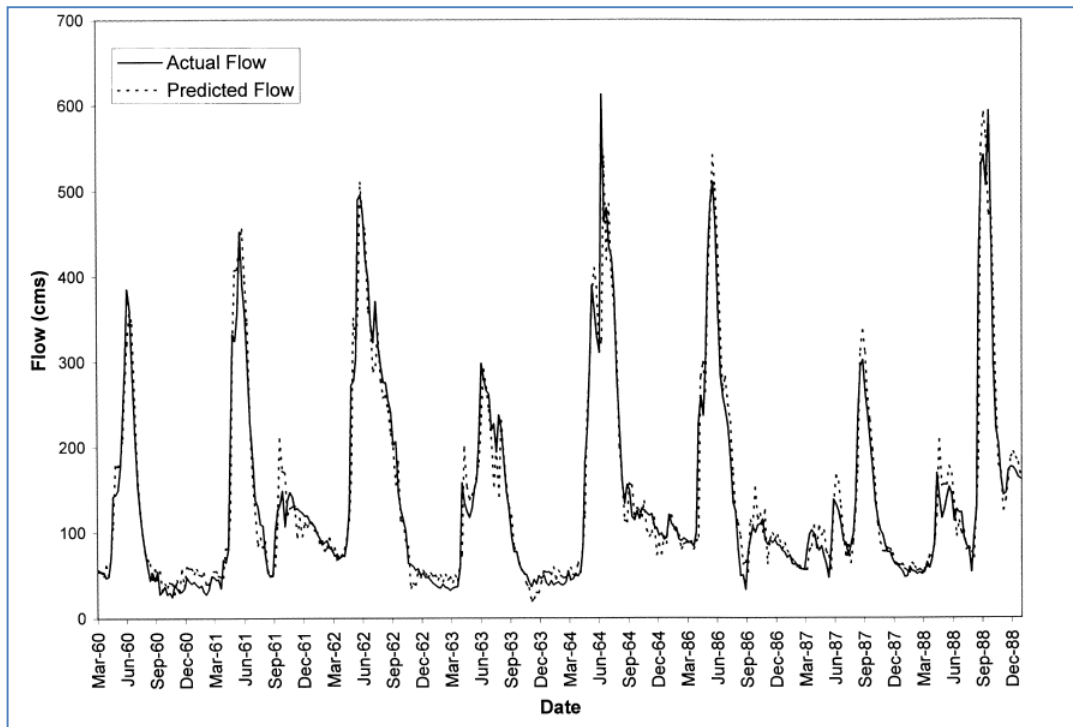
Figure 2-4: One Week Ahead Forecast of Flow into Namakan Lake
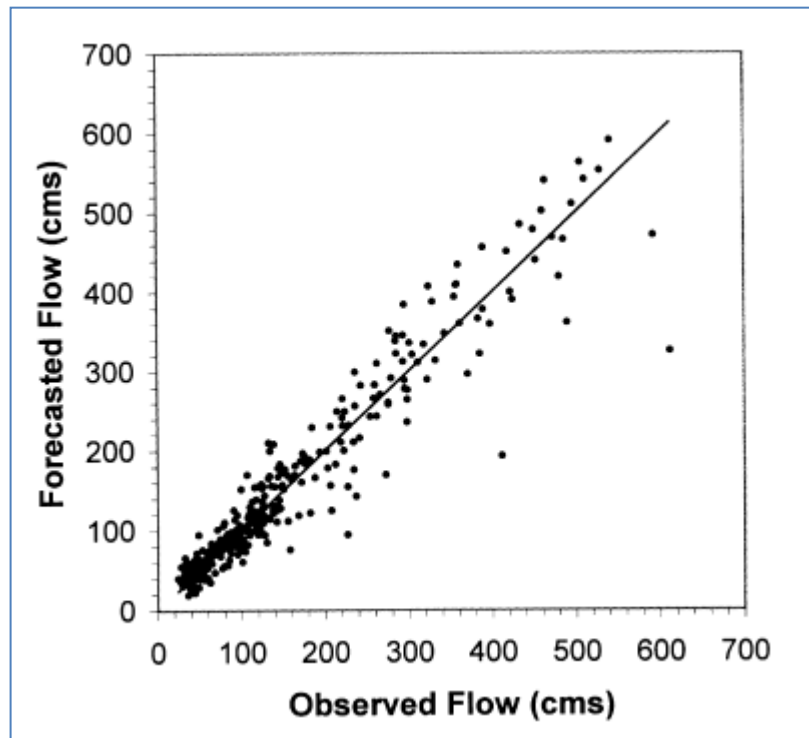
Figure 2-5: Observed vs. Forecast of Flow into Namakan Lake

2.3.3 Advantages of ANN

Since few years back, research regarding to neural network is slightly extensive growth is wide in gamut of scientific research field. [11]. In the context of automatic control theory neural networks can he characterized by the:-

- neural networks is one of the reliable predictor for arbitrary nonlinear mappings;
- the character of parallel structure able to increase the stability in the network and the speed of information processing;
- For real time control, neural networks can be performed effectively quick after the training is completed;
- With help of the quantitative information, neural networks can be applied in order to solve with qualitative cases as well;
- neural networks are naturally applied to multidimensional systems.

2.4 Least Square Vector Regression

ANN review has been gone through and now the second method of intelligent load forecasting shall be discussed in this section. Suykens et al. had introduced the least squares support vector machine (LS-SVR) as generalization and simplifiers to standard SVR because instead of traditional quadratic programming method a consideration into a great extent by applying linear least squares criteria to the error function [12]. LS-SVR carries forward the advantages of SVR in many angles. Some of that is able to generalize remarkably and a unique solution promotes the applications of LS-SVR in many pattern recognition and regression problems. In SVR, the first step of the process is when regression problem is formulated and represented as a convex quadratic programming (QQP) problem which is one of a special case involving general field of convex optimization. Then, the SVR regressor basically plots the inputs into a higher dimensional feature space in which a linear regressor is assembled by minimizing an appropriate cost function [13]. By solving a finite dimensional QP problem, the

regressor is defined according to Mercer's theorem. The purpose of it is to avoid explicit knowledge of the high dimensional mapping and using only the related kernel function.

2.4.1 Least Square Vector Regression Architecture

To mention about data classification and regression, LS-SVM is a one of the reliable technique [21, 22]. According to Y. Li, T. Fang and E. Yu, given a training data set of N points with input data where ξ and ξ^* are the slack variables and C is the regularization parameter. [13];

$$\Psi(\omega, \xi, \xi^*) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

(2.5)

$$\text{Where} \begin{cases} y_i - \omega^T\varphi(x_i) - b \leq \varepsilon + \xi_i \\ \omega^T\varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi, \xi^* \geq 0, \ i = 1,2,\dots,l \end{cases}$$

(2.6)

LS-SVR functioning by applying the linear least square criteria to the lost function compared with standard SVR, LS-SVR instead of inequality constraints with equality ones. LS-SVR can be illustrated by [13];

$$\Psi(\omega, e) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}e_i^2$$

(2.7)

15

Subject to the equality constrains

$$y_i = \omega^T \varphi(x_i) + b + e_i, i = 1,2, \dots, l$$

(2.8)

The Lagrangian function can be constructed in order to modify the standard SVR into LS-SVR by

$$L(\omega, b, e, \alpha) = \Psi(\omega, e) - \sum_{i=1}^{l} \alpha_i (\omega^T \varphi(x_i) + b + e_i - y_i)$$

(2.9)

Where α_i $(i = 1, 2, \dots, l)$ are Langrange multipliers, it is related to Karush–Kuhn–Tucker (KKT) conditions for optimality [13], given by

$$
\begin{cases}
\dfrac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{i=1}^{l} \alpha_i \varphi(x_i) \\[2mm]
\dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i = 0 \\[2mm]
\dfrac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = C e_i, i = 1,2, \dots, i \\[2mm]
\dfrac{\partial L}{\partial \alpha_i} = 0 \rightarrow \omega^T \varphi(x_i) + b + e_i - y_i = 0, i = 1,2, \dots, i
\end{cases}
$$

(2.10)

After elimination of the variables ω and e_i , the optimization problem can be transformed into;

$$
\begin{bmatrix} 0 & 1_l^T \\ 1_l & \Omega + C^{-1}l \end{bmatrix}
\begin{bmatrix} b \\ \alpha \end{bmatrix} =
\begin{bmatrix} 0 \\ y \end{bmatrix}
$$

16

Where, $1_l = [1, ..., 1]^T, \alpha = [\alpha_1, ..., \alpha_l]^T$ and $y = [y_1, ..., y_l]^T$. The Mercer's condition is applied within the matrix $\Omega_{ij} = \{\Omega_{ij}\}_{l \times l}$ as follows

$$\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$$

LS-SVM model for regression becomes

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x_j) + b$$

where α_i and b are the solution to the linear Eq. (2.11). Then, evaluation upon the error analysis is will be done for this forecasting technique hence Mean Absolute Percentage Error (MAPE), ε is the formulae for after-the-fact error analysis:

$$\varepsilon = \frac{1}{l} \sum_{i=1}^{l} \frac{|f(x_i) - y_i|}{y_i} \times 100$$

where $f(x_i)$ is the actual load and y_i is the forecasted load [13].

In previous chapter, the procedure of ANN has been discussed in order to maximize comprehension upon the flow of progress of ANN. Now, the procedure of LS-SVR will be elaborate briefly in diagram below:
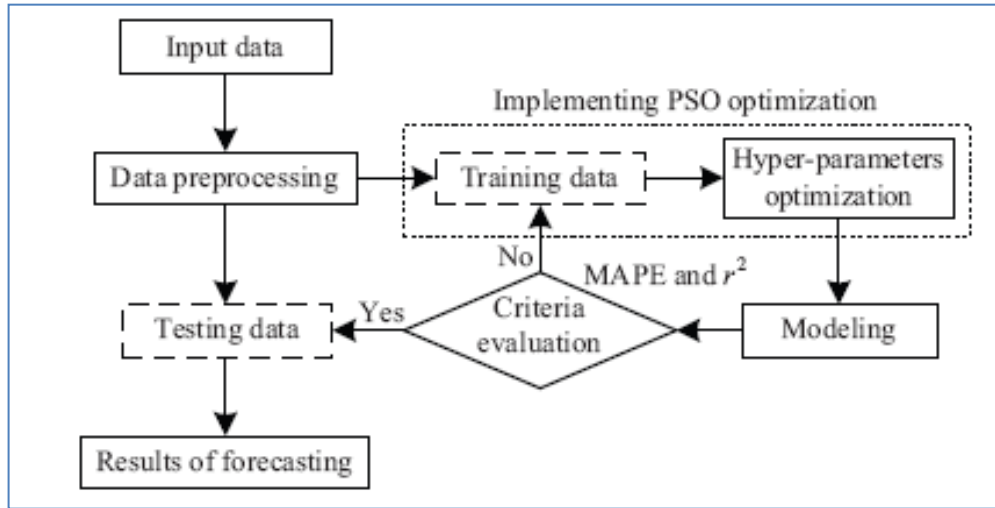
Figure 2-6: Diagram of Procedure Structure

2.4.2 Application Examples

By referring to some examples on application of LS-SVR, there will be some review on the selected paper which has high citation. The selection paper will be based on the objective itself which is the beneficial of the implementation compare to ANN in load forecasting in various fields. Those papers are listed as follow:

2.4.2.1 On-line least squares support vector machine algorithm in gas prediction [14]

The safety data set comes from a coal mine that has a thick gas coal bed in Shanxi Province. According to the analysis of the production arrangement and the safety data, it is essential for peak production periods of gas emission at times between 4:00–5:00 and 8:00–9:00. The data of the first twenty days of November and December in 2007 were used to carry out the training. The on-line learning LS-SVR prediction algorithm was used to predict the data of the last ten days of November and December. Gas data were sampled every two minutes, with 30 data sets per hour.

18

Result:

A new on-line learning LS-SVR prediction algorithm has been proposed in order to replace the traditional off-line prediction algorithms of coal mine safety which is without a dynamic prediction function. Actual gas emissions in a coal mine were predicted with this new algorithm. The algorithm has proven to be effective by comparing the observed data with other, related algorithms.
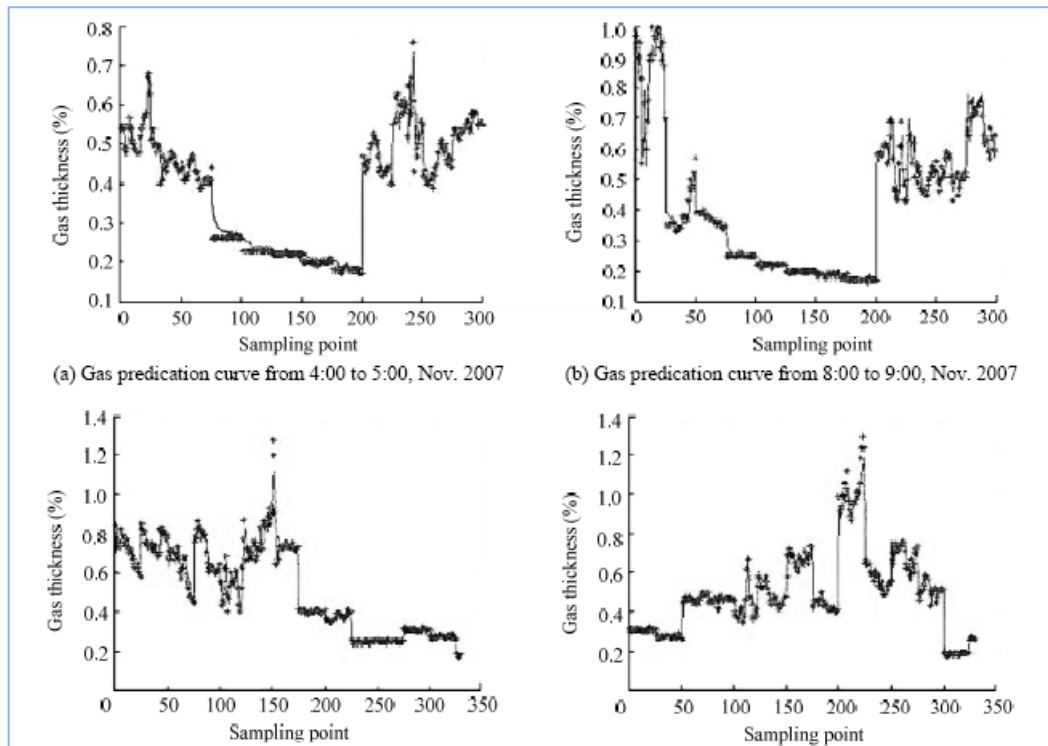
Figure 2-7: Gas Prediction result using Ls-SVR

2.4.3 Advantages of LS-SVR

Comparing with ANN, LS-SVR has its own advantages in order to be consider as one compatible option as the forecasting model for capture missing data of the gas pipeline although the main disadvantage is lack of sparseness of solutions. The pros of LS-SVR are listed as below:

- LS-SVR is computationally very cheap in cost but it still operates in the same important properties of the SVM [15].
- Only few tuning parameters required in order to proceed with LS-SVM formulation [16].
- LS-SVR provides the efficient prediction of additive models with respect to existing practice, but in the same time able to interpret of the estimated model, opportunities towards structure detection and the connection with existing statistical techniques [17].
- Support vector regression (SVR) is also remarkable to train continuous and non linear data which is suitable for a longer process. [17]

Genetic Algorithm method will be discussed in section below.

2.5   Genetic Algorithm

Along the globalization of science and technology, the principle genetic algorithm (GA) is evolve in biology to computer science and due to that circumstances, GA is usually being proposed to find optimize the approximation of solutions [1]. By identify from one population of points to another, GA is a procedure that continuously samples the parameter space, the search is directed toward the area of the best solution so far. This algorithm is the most outstanding perform in obtaining global solutions for non-linear function of neural network compare to backpropogation [18]. There are three most important aspects should be consider of using GA including, definition of the objective function, definition and implementation of the genetic representation and definition and implementation of genetic operators [1].

2.5.1   Artificial Neural Network Combined with Genetic Algorithm

The function of GA is not to training back the prediction procedure but to optimize the performance made by ANN. GA is commonly used for combination of two algorithms through the application of the GA to one particularly complex intelligence function. It has been proved that GA can provide better results for training a feedforward neural network than the traditional techniques of backpropagation [18].

As we can see, ANN naturally was trained to evoke to itself or predict its input vector. Assume $\vec{Y}$ is the ANN output vector, hence an ANN formula is [1]:

$$\vec{Y} = f(\vec{X}, \overrightarrow{W})$$

$$(2.15)$$

$\vec{X}$ is the input vector while $\overrightarrow{W}$ the vector of weights. As stated before that ANN can predict its own vector, approximation can be made whereby vector $\vec{X}$ is equal to $\vec{Y}$. As

we can see, actually in real time series, the input vector $\vec{X}$ and output vector $\vec{Y}$ will not always be perfectly the same hence, we will have an error function expressed as the difference between the input and output vector.

Thus, the error can be formulated as

$$e = \vec{X} - \vec{Y}$$

(2.16)

By substituting equation (2.15) into (2.16)

$$e = \vec{X} - f(\vec{X}, \vec{W})$$

(2.17)

In order to minimize the error and extract non-negative values, the error function can be rewritten as the square of equation (2.17)

$$e = (\vec{X} - f(\vec{X}, \vec{W}))^2$$

(2.18)

To relate with the problem statement of this project, some of the values for the input vector $\vec{X}$ are not available due to missing data by the metering system. Hence, we can categorize the input vector $(\vec{X})$ elements into $\vec{X}$ known represented by $(\vec{X_k})$ and $\vec{X}$ unknown represented by $(\vec{X_u})$. Rewriting Equation (2.17)

$$e = \left(\left\{ \begin{matrix} \vec{X_k} \\ \vec{X_u} \end{matrix} \right\} - f\left(\left\{ \begin{matrix} \vec{X_k} \\ \vec{X_u} \end{matrix} \right\}, \vec{W} \right) \right)^2$$

(2.19)

Equation (2.19) is minimized using GA to approximate the missing input a value because it finds the global optimum solution [18]. Since a genetic algorithm always

finds the maximum value, the negative of Equation (2.19) was supplied to the GA as a fitness function. Thus, the final error function minimized using the genetic algorithm is

$$e = -\left(\left\{\frac{\overrightarrow{X_k}}{\overrightarrow{X_u}}\right\} - f\left(\left\{\frac{\overrightarrow{X_k}}{\overrightarrow{X_u}}\right\}, \overrightarrow{W}\right)\right)^2$$

(2.20)

Figure below illustrated the schematic representation of proposed model. Started with the simulation of neural network, the error function is defined from the inputs and output vector obtained. The error function is then being minimized when genetic algorithm recall the error function from neural network regarding to the Equation (2.19) to approximate the missing variables in the error function [1].
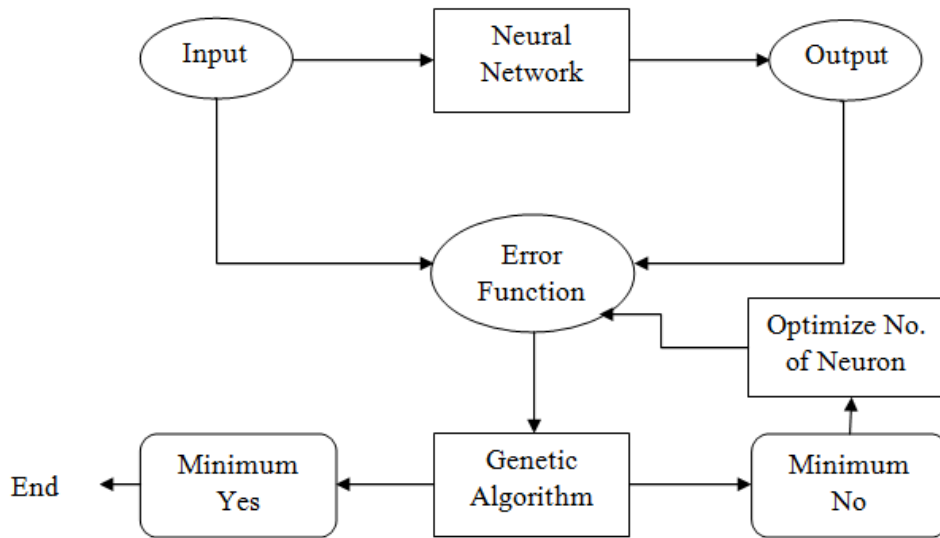
Figure 2-8: Schematic Representation of ANN combined with GA

2.5.2 Least Square Support Vector Regression Combined with Genetic Algorithm

GA also can be combined with LS-SVR in order to predict the missing data of the gas pipeline. As we discussed before, the specialty and specific operation of Support vector Regressor (SVR) is to solve nonlinear regression and time series problems [19]. However, the application for load forecasting is rare. In order to utilize GA in LS-SVR, GA will determine free parameters of support vector machines.

As we discussed in previous section, it is stated that the LS-SVR function is [19]:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x_j) + b$$

(2.21)

Any function that meets Mercer's condition [19, 20] can be used as the Kernel function. Although LS-SVR is the more compatible in order to determine the regression factor, still the method for confirming efficiently the selection of parameters is not as reliable as we aim for. Therefore, GA is a big rule in order to propose LS-SVR model to optimize parameter selection. The architecture of a LS-SVR model is being illustrated as below by passing it from generation to generation. From LS-SVR forecasting the error function, MAPE will be determine whether it achieve the required minimum error or not. The error function that did not achieve the requirement will pass to GA and the parameters selection will be forward to LS-SVR back to determine the error until parameters being obtained.

LS-SVR with GA

GAs
- Parameter selection
- No
- Stop condition

LS-SVR
- Forecasting
- Error calculation

Yes — Parameters obtained
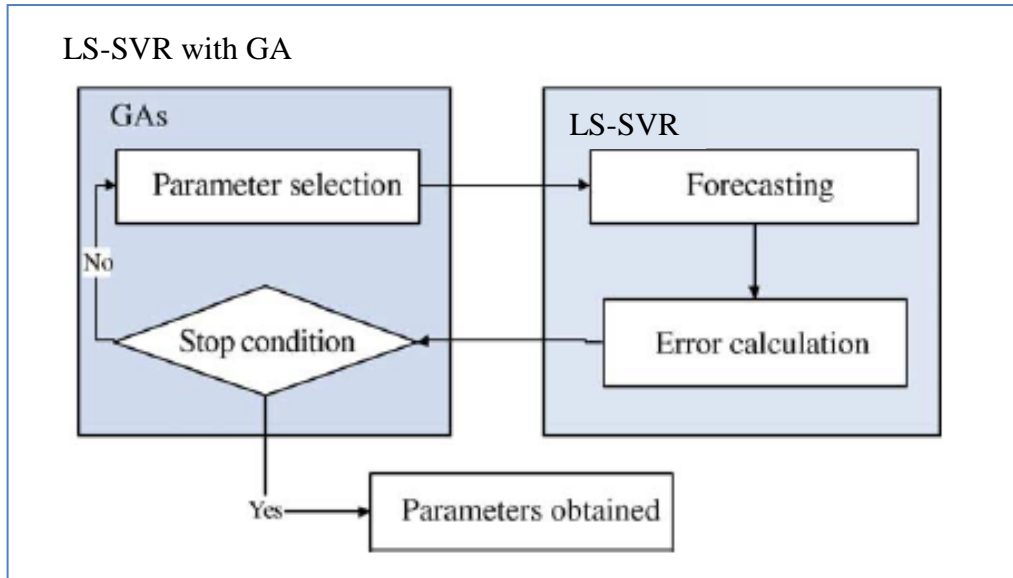
Figure 2-9: Architecture of LS-SVR with GA model

2.5.3   Advantages of Genetic Algorithm

To select the best model for gas consumption forecasting technique, the performance comparison of effectiveness, ease-of-use, and efficiency of the methods being analyze. Comparison between GA combined with ANN and GA with LS-SVR is as below [1,19]:

Table 1: Advantages of Combination with GA

ANN + GA	LS-SVR + GA
• GA will estimate the true functional form by refers to the accuracy of each algorithm • GA with ANN is much more simpler to simulate and learning process is not going to take so long • The CPU time taken by each algorithm for converging upon the best found solutions in training ANN • The optimal architecture, connectivity and training parameter for neural network can be defined easily by GA that calling back the parameter function from ANN • GA behavior of mimic biological evolution and progressively seek superior solutions using the guide of selective function (fitness) and partial combination of candidate solutions.	• This model has nonlinear mapping capabilities and thus can more easily capture data patterns than can the ANN and regression models. • The superior optimization of GA contribute to the efficient performance of the advantage of LS-SVR that predict well than ANN • GA and LS-SVR model performs structural risk minimization rather than ANN that minimizing the training errors. • it converges very fast to the optimum, especially when the initial guess is close to the optimum • The computation time is lower than the other classifier algorithms

2.6   Selection of Prediction Model

As far as went through the literature review, by comparing all four methods, ANN combining with GA is the best solution for this project. This is because of various factor especially time limit for completion of the project. The limitation of duration for both FYP 1 and FYP 2, development of ANN and GA is much shorter compare to LSSVR and GA. Besides, the advantages of ANN and GA which is less time consuming and also less deals the effort of finding optimal algorithm can be considered. GA also have the ability to estimate the true functional form by refers to the accuracy of each algorithm. The facilities and utilities also contribute to the selection that has been made. By combining both algorithms, the coding development can be done using MATLAB software that is user friendly.

For next section, we will focus on methodology including data selecting and filtration, algorithm selection and model development that contribute to the completion of the project.

CHAPTER 3

METHODOLOGY

3.1   Research Methodology and Project Activities

For this Final Year Project, the research methodology is separated into five main stages and the activities planned for each stage are explained one by one. The first stage is the preliminary research work on the project including preparation of proposal. Then, the development of the missing data prediction model will be constructed using the MATLAB software. On the third stage, testing and validation work of the model will be done followed by the results analysis and discussion. The last stage is the final documentation that compiles all the research works and the outcomes of the project.

3.1.1   Preliminary Research Work

This stage involves compilation and review of information from various books, journals and technical papers on related topics. Important information required for the project such as categories of prediction models, comparison will be done to select which is the most preferable model of prediction. This information is to be used to conduct the literature review and for next stages of the project and also to enhance the knowledge upon related topic. Instructions and tutorials on MATLAB software are also reviewed.

3.1.2   Model Development

Raw data which are pressure (P), temperature (T), gross volume (Vg), calorific value (CV) and specific gravity (sg) for the project will be collected from the actual plant which in this case is from PETRONAS Gas Bhd. and sorted to its input and output for

model development. The energy prediction model will be developed by building neural network block using Simulink in the MATLAB software. A few models will be simulated in the software.
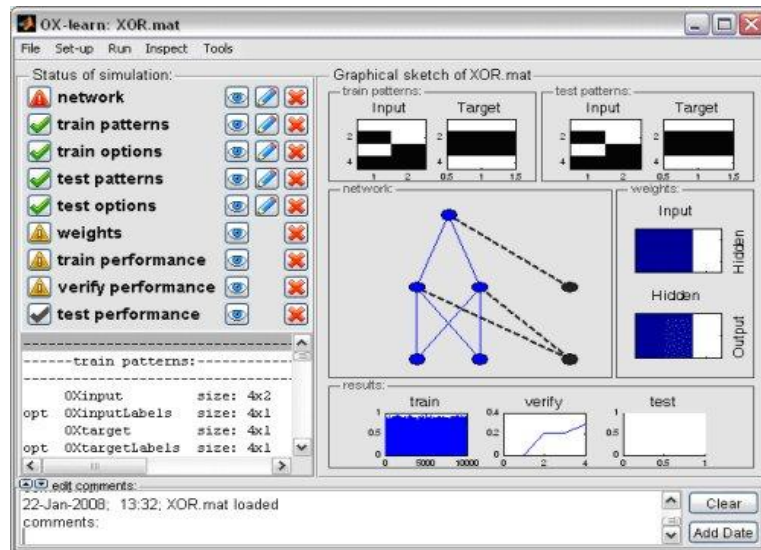
Figure 3-1: Example of MATLAB Application
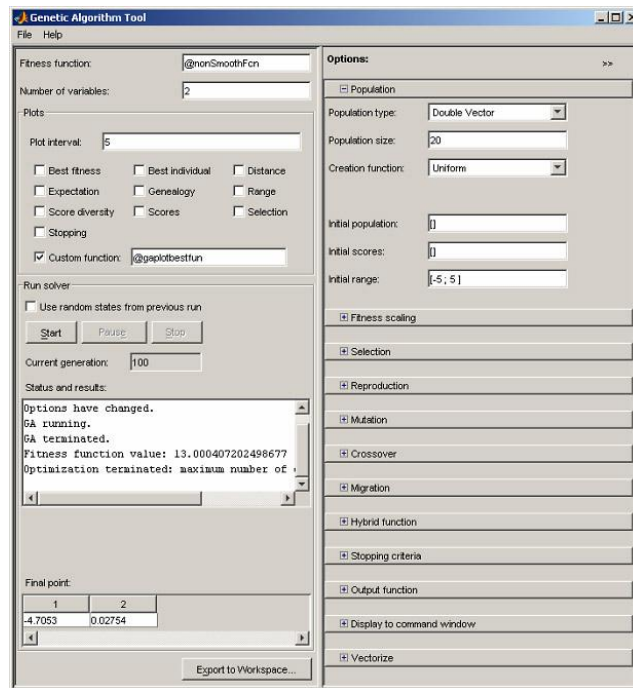
Figure 3-2: Example of MATLAB Application

3.1.3 Testing and Validation Work

The models that have been developed from the previous stage will be tested and validated for its efficiency and accuracy. A successful model is the one that meet all the objectives listed. The best model will be chosen based on the performance indicator. The successful model will proceed to the next stage while the unsuccessful ones will be tested and validated again. But during this term, various changes of modification also will be done parallel with the progress of the project.

3.1.4 Results Analysis and Discussion

The simulation results will be analysed and discussed with the supervisor and experts. The model will be evaluated to make sure it meets the requirements of the project. The simulation will be re-run for further investigation to improve the results if necessary. All data collected will be used in order to defense the reliability of the project.

3.1.5 Final Documentation

The research works, methods and outcomes of the project will be documented for future use. All of the documents required for the project will be compiled continuously starting at the early stage of the project until the last stage so as to keep track on the project and not to wait until the last minute. Final documentation will involved all stages including early assumption and modification being done.

3.1

Start

Preliminary Research Work

Model Development

Testing & Validation Work

No

Successful

Yes

Result Analysis & Discussion

Improvement

Yes

No

Final Documentation

End
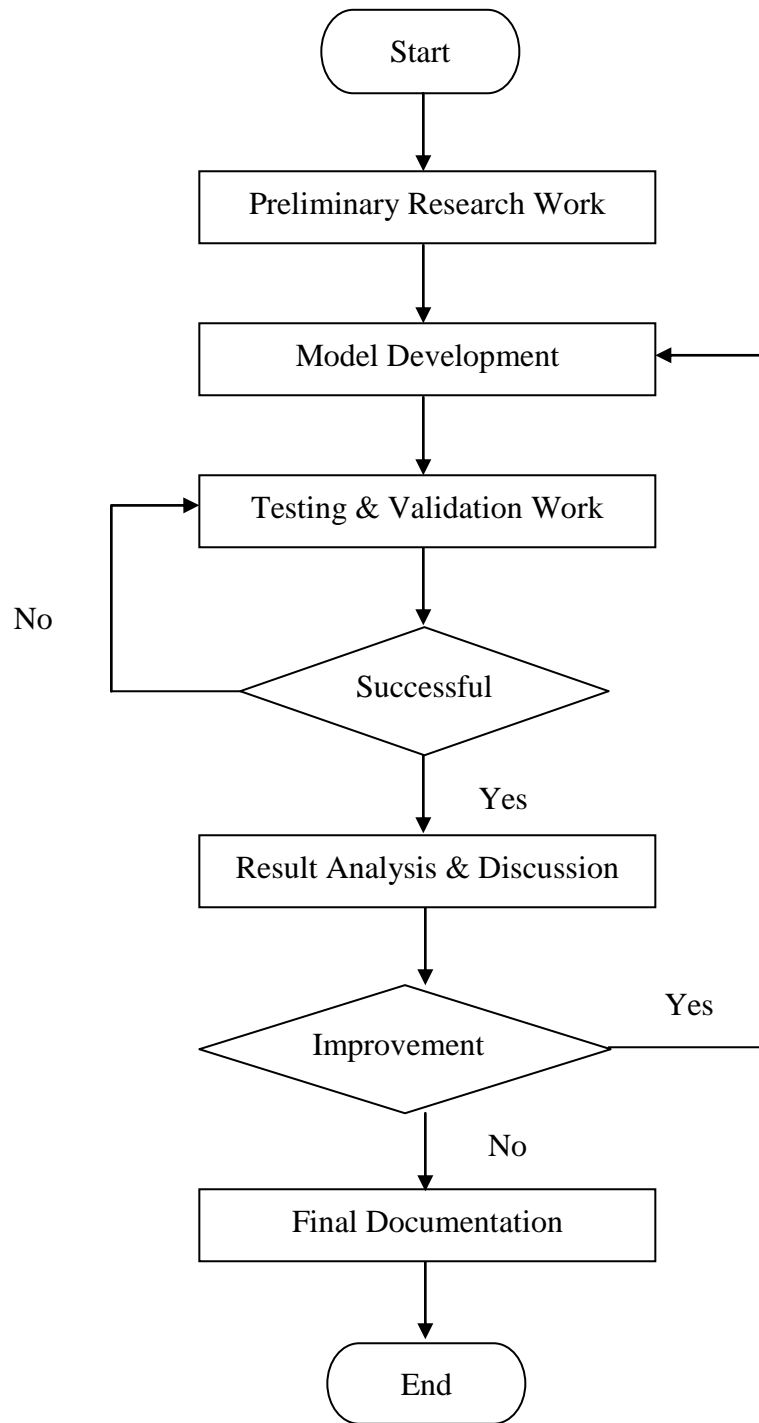
Figure 3-3: Flowchart of Methodology

3.2 Project Activities

At this point, in order to achieve the reliability of the project, activities must obey the schedule based on development of forecasting is as shown in Figure 3-3. It divided into five main parts which are; collection and filtration of data, research and selection of network architecture, determination of learning algorithm and activation function, determination of training and validation data division and determination of number of neurons. All section will be discussed deeply in the next section.
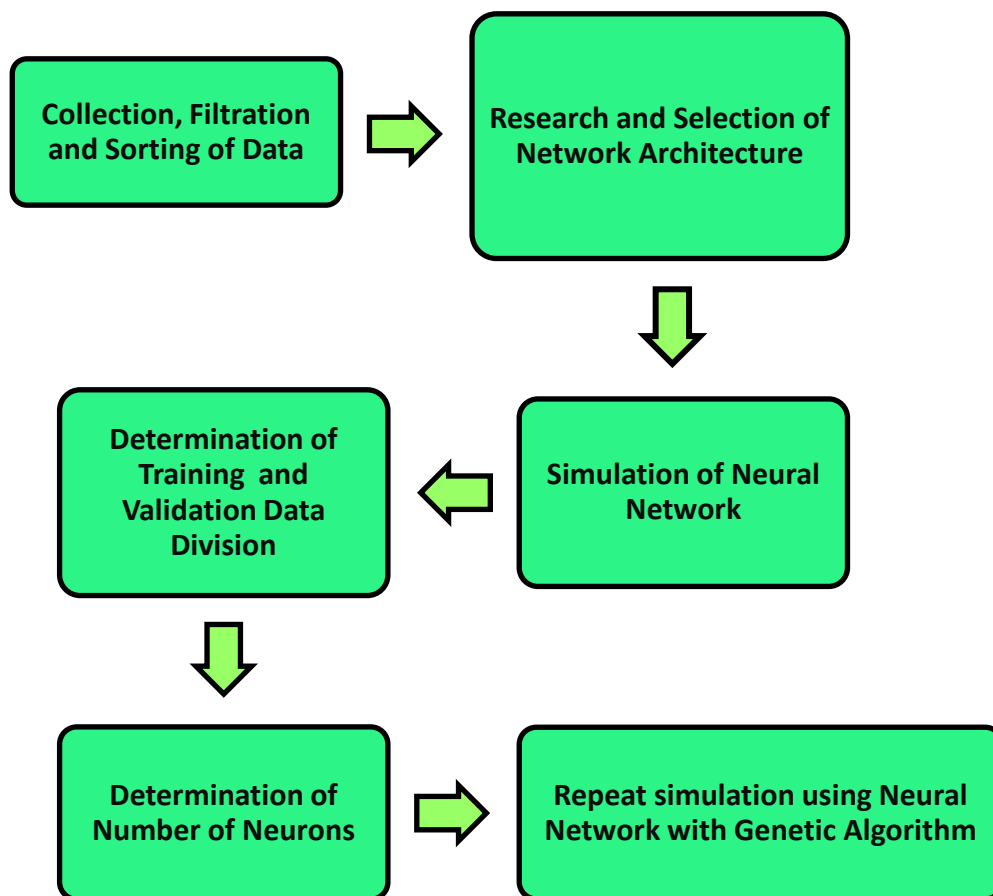
Figure 3-4: Development of Forecasting Model Flowchart

3.2.1 Collection and Filtering of Data

For training and validation of the model purposes, raw data are collected from Nur Metering Station. The data are taken for six-month duration on hourly basis every day from the flow computer and gas chromatography in the metering system. These data are sorted according to months, peak hour usage and non-peak hour usage; and all invalid and irrelevant data are filtered out. Data filtered are data that are out of ranges or abnormal such as spike or zero reading. There are 3442 data with an average of 574 every month after the filtering process. Data of six months (January – June) consecutively are trained and validated for the model. The filtered data are then sorted to its input and output.
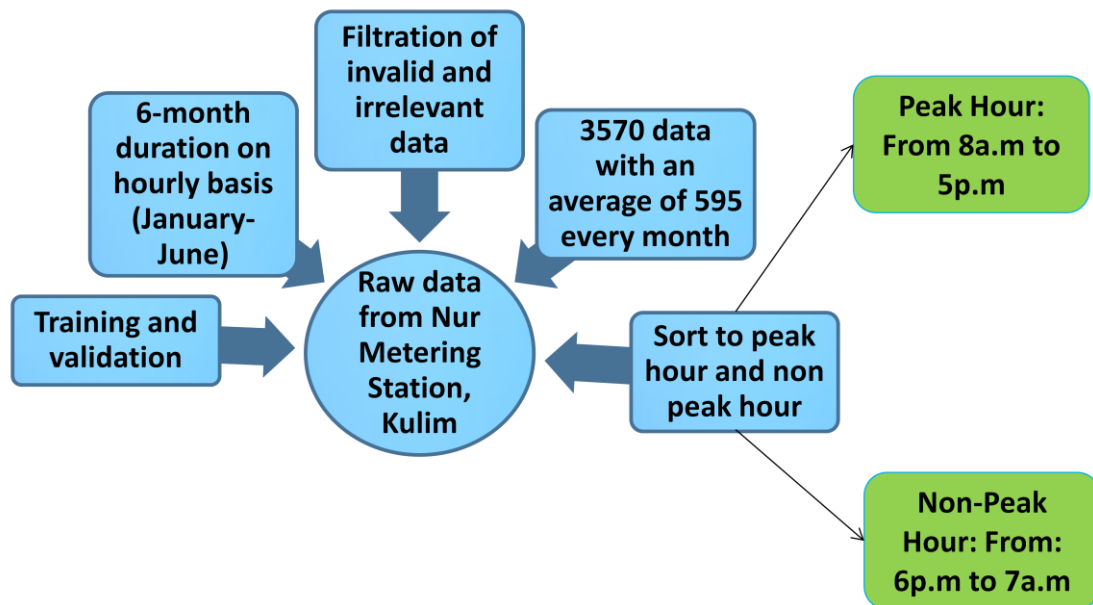
Figure 3-5: Collection and Filtration of Data

3.2.2   Research and Selection of Network Architecture

Basically, for this project, we purposely focus on development of ANN architecture during FYP 1, whereby it will be optimized by GA during FYP2. As we can see the research and determination of network architecture has been done by J. Paris [10] while this paper is responsible in construction of combination with GA. Generally, there are three types of fundamentals network architectures that are single-layer feedforward network, multilayer feedforward network also known as multilayer perceptron and recurrent network. The multilayer feedforward network is the most popular network architecture and it has been selected for the development of the neural network energy prediction model. According to a research by Ismail and Ibrahim [24], multilayer feedforward layer is suggested for the development of the model as it is architecture with most suitable and flexible network. Three layers which are input layer, hidden layer and output layer are being constructed to complete the structure of the architecture.

3.2.3   Simulation of Neural Network

The neural network is trained using Levenberg-Marquardt with Bayesian regularization algorithm (trainbr) [23]. In this investigation, the epochs had been made constant up to 1000 in order to determine the suitable neurons and set of data that need to be used in this model. The value of initial weight had been set randomly because it did not affect much of the simulation process.

The role of activation functions for the hidden layer and output layer is to scale data output from a layer. To complete the project, the activation function that is suitable for this case is tan-sigmoid for the hidden layer and output layer. The selection of activation function analysis has been done by J. Paris in previous session and she concluded that, Levenberg-Marquardt algorithm with Bayesian Regularization is the best learning algorithm based on its percentage of error with activation function of tan-sigmoid for both hidden layer and output layer. The percentage of error for training phase is up to

4.7695% while for validation is 2.9919%. From the result of her investigation extension of performance optimization being done in order to achieve the objective of this project.

In order to calculate the difference between actual and prediction data, root mean square (RMSE) is used as a performance indicator. The randomness or because the estimator does not account for information is some of the factor that contributed to the existence of difference between actual and prediction energy consumption. The least RMSE of the model will be selected as the preferable model to be use for the next evaluation.

The equation of RMSE is:-

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_p - y)^z}$$

(3.1)

Where (y_p) is the predicted energy and (y) is the current energy for each training data set and validation data set whereby the number of data is N.

To further justify the performance of the model, the difference occurs between prediction model and actual data also being defined using percentage of error. Logically, the lower the percentage of error, the more accurate the model is. The equation for percentage of error is:

$$Percentage\ of\ error = \frac{y_p - y}{y} \times 100\%$$

(3.2)

3.2.5 Determination of Number of Neurons

The preferable number of neurons is selected base on the one that can generate the least RMSE value and percentage error for both training and validation phase. Certain parameters are made constant which are the learning algorithm using the Levenberg-Marquardt algorithm with Bayesian Regularization and the activation functions tan-sigmoid for the hidden layer and pure-linear for the output layer. The neural network is tested with one neuron up to the number that gives the least value of RMSE and also tested slightly more than the optimum number of neurons.

3.2.6 Repeat simulation using Neural Network with Genetic Algorithm

This part is where the ANN is being optimized by GA as being discussed in the literature review part. When ANN had predicted the energy consumption and error between the predictions had been compared,

$$e = \vec{y_p} - \vec{y}$$

$$(3.3)$$

In order to minimize the error and extract non-negative values, the error function can be rewritten as the square of Equation (3.3)

$$e = (\vec{y_p} - \vec{y})^2$$

$$(3.4)$$

Since a genetic algorithm always finds the maximum value, the negative of Equation (3.4) was supplied to the GA as a fitness function. Thus, the final error function minimized using the genetic algorithm is

$$e = -(\vec{y_p} - \vec{y})^2$$

$$(3.5)$$

In order to complete the GA part, the MATLAB implementation of genetic algorithm described in Houck et al. [1995] has been used. The genetic operator values such as number of generation, mutation, crossover, etc. was being altered to evaluate genetic algorithm.

After all the procedure had been done, it is essential to discuss and analyze the result obtained. This task will be performed in the next section.

3.3   Tool Required

The tool involved in this project only software tools. It is required for the modeling and simulation process of the project is MATLAB software. MATLAB software is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. MATLAB is the most common software being used because of the user friendly characteristic and the software has built in ready-to-use functions that have been optimized for fast execution and easy access to toolboxes and user generated contributions that are useful for quick implementation and test for various approaches before committing to a single one during the research and development process. Its graphical user interface (GUI) and display functions can help visualize data without spending hours coding in more complex languages. Furthermore, MATLAB will give the outcome graphically hence analysis and discussion can be done vividly. These aspects have made the MATLAB software as the most suitable tool for this project.

3.4 Gantt Chart

The Gantt Charts in Table 1 and 2 show the timeline of the project for two semesters. The starting and finishing weeks for all activities are illustrated. The schedule for each activity can change from time to time depending on the work progress.

3.3.1 Gantt Chart: First Semester

The Gantt Chart for the first semester is as shown below.

Table 2: Gantt Chart for first semester

No.	Detail/Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Selection and Confirmation of Project Title	█													
2	Preliminary Research on Topic		█	█	█	█									
3	Submission of Extended Proposal						●								
4	Study on MATLAB for Model Development						█	█	█						
5	Proposal Defence								█	█					
6	Model Development in MATLAB									█	█	█	█		
7	Submission of Interim Draft Report													●	
8	Submission of Interim Final Report														●

38

3.3.2 Gantt Chart: Second Semester

The Gantt Chart for the second semester is as shown below.

Table 3: Gantt Chart for Second Semester

No.	Detail/Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Model Development and Modification	▨	▨	▨	▨	▨	▨								
2	Testing and Validation							▨	▨						
3	Submission of Progress Report								●						
4	Result Analysis and Discussion								▨	▨	▨				
5	Pre-EDX										●				
6	Submission of Draft Report											●			
7	Submission of Dissertation (soft bound)												●		
8	Submission of Technical Paper												●		
9	Oral Presentation													●	
10	Submission of Project Dissertation														●

3.6 Key Milestone

A key milestone is constructed to mark the end stage of a work or process of the project. It is an important element in order to monitor the progress and make sure that the project is on schedule. The key milestone for both first and second semesters is as shown in Table 4 below.

Table 4: Key milestone for first semester and second semester

	FIRST SEMESTER													
No.	**Detail/ Week**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	Completion of preliminary research work								■						
2	Completion of Interim Report (Draft and Final)													■	
	SECOND SEMESTER														
No.	**Detail/ Week**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
3	Completion of model development						■								
4	Completion of testing and validation work								■						
5	Completion of results analysis and discussion										■				
6	Completion of final documentation												■		

CHAPTER 4

RESULT AND DISCUSSION

4.1 Chapter Overview

This chapter accumulates the results and discussions of the project along Final Year Project phase. In continuity of the project, it has been decided to use combination of ANN with GA as priority prediction model and will be compare to ANN only prediction model. This section contains the results on parameters selection for the model such as the network architecture, learning algorithm, activation function, training and validation data division and number of neurons in ANN architecture without being combined with GA. Each result are then analyzed and discussed later the project will be proceeding with model that embraces the usage of ANN that being optimized with GA.

4.2 Data Gathering and Analysis

The data used for the development of neural network model is the actual historical data taken from the metering system flow computer (FC) in Transmission Operation Division (TOD), PETRONAS Gas Berhad (PGB), Gurun. The data consists of the actual energy data obtained from FC and value of input data from all the measuring equipments. As stated before, data will be filter and sort according to peak hour duration and also non peak hour duration of gas consumption. As for peak hour usage duration it is consist from 0800 hour until 1800 hour and for non-peak hour is from 1900 hour until 0700 hour.

Month	Week	Day	Hour	Temperature	Pressure	CV	Sg	Volume	Energy
			8	27.38	3291.67	36.79	0.68	811.1	1029.1
			9	27.51	3331.23	36.75	0.68	805.5	1033.1
			10	27.94	3367.43	36.73	0.68	808.1	1045.7
			11	28.34	3396.86	36.82	0.68	806.7	1054.5
			12	28.42	3426.25	36.75	0.68	804.1	1058.0
		1	13	28.74	3458.22	36.74	0.68	791.6	1050.1
			14	29.10	3490.64	36.68	0.68	788.7	1053.2
			15	28.80	3523.34	36.68	0.68	779.0	1051.8
			16	28.16	3557.26	36.68	0.68	772.1	1056.1
			17	28.02	3593.01	36.73	0.68	762.7	1056.4
			18	27.85	3627.44	36.77	0.68	748.4	1048.9
			8	27.83	4435.43	36.80	0.68	668.3	1161.0
			9	27.90	4489.43	36.82	0.68	667.6	1175.3
			10	28.15	4544.35	36.83	0.68	680.4	1212.6
			11	28.57	4602.12	36.82	0.68	683.4	1231.7
			12	28.95	4657.80	36.80	0.68	710.6	1294.7
		2	13	29.12	4715.63	36.75	0.68	755.4	1391.3
			14	29.22	4777.90	36.78	0.68	751.7	1405.1
			15	29.35	4838.66	36.77	0.68	748.9	1417.4
			16	29.39	4898.41	36.84	0.68	748.7	1438.8
			17	29.07	4957.35	36.85	0.68	745.0	1453.2
			18	28.71	5013.86	36.82	0.68	734.1	1450.5
			8	27.95	5633.18	36.70	0.68	643.8	1447.0
			9	27.93	5660.67	36.71	0.68	657.8	1487.3

Figure 4-1: The actual data from Gurun Metering Station Daily Report (peak hour)

Month	Week	Day	Hour	Temperature	Pressure	CV	Sg	Volume	Energy
			7	27.37	3253.30	36.81	0.68	830.8	1041.9
			19	27.45	3666.36	36.78	0.68	729.8	1036.2
			20	27.62	3710.51	36.77	0.68	728.5	1046.6
			21	27.47	3762.25	36.75	0.68	719.5	1049.0
			22	27.61	3819.41	36.72	0.68	722.5	1068.5
			23	27.68	3877.60	36.67	0.68	723.0	1084.9
		1	0	27.70	3938.31	36.73	0.68	717.0	1095.6
			1	27.72	3998.41	36.76	0.68	707.5	1099.4
			2	27.71	4059.14	36.79	0.68	709.8	1122.0
			3	27.77	4119.98	36.78	0.68	726.6	1166.1
			4	27.80	4182.31	36.74	0.68	731.4	1191.2
			5	27.79	4246.12	36.72	0.68	694.1	1148.4
			6	27.80	4311.16	36.69	0.68	689.9	1159.2
			7	27.82	4375.24	36.73	0.68	679.2	1160.5
			19	28.44	5067.54	36.81	0.68	723.2	1447.8
			20	28.33	5114.02	36.77	0.68	724.1	1462.9
			21	28.21	5151.52	36.73	0.68	720.3	1466.5
			22	28.13	5179.27	36.63	0.68	713.7	1458.0
			23	28.08	5207.19	36.62	0.68	710.7	1460.9
		2	0	28.05	5239.99	36.68	0.68	701.2	1454.2
			1	28.03	5282.25	36.70	0.68	688.5	1441.3
			2	28.02	5329.45	36.71	0.68	679.1	1435.9
			3	28.00	5378.29	36.70	0.68	672.4	1435.5
			4	28.02	5429.56	36.65	0.68	665.2	1433.3

Figure 4-2: The actual data from Gurun Metering Station Daily Report (non-peak hour)

4.3   Simulation of Neural Network

The learning algorithm used is Levenberg-Marquardt algorithm with Bayesian regularization with activation function of tan-sigmoid for hidden layer and pure-linear for output layer. The number of neurons that gives the lowest MSE value and percentage of error for both training and validation data is determined by using Levenberg-Marquardt with Bayesian Regularization as its learning algorithm. The activation functions used are tan-sigmoid for hidden layer and output layer. The number of neurons is manipulated in the hidden layer from 1 neuron up to the number that gives the least value of RMSE and also tested slightly more than the optimum number of neurons. In this case, the model is tested from 1 neuron to 12 neurons. The number of neurons in the output layer is made constant to 1 neuron. The purpose of selecting the range is because to be compatible with the values of data which 1989 is for peak hour duration and 2335 for non-peak hour duration. In addition, Levenberg-Marquardt algorithm with Bayesian Regularization is selected as the best learning algorithm based on its percentage of error with activation function of tan-sigmoid for hidden layer and pure-linear for output layer.

Table 5: RMSE and % Error for Neural Network Only

Model and Data Division	No. of Neuron	RMSE (GJ)		Percentage of Error	
		Training	Validation	Training	Validation
	1	107.3356	93.2945	3.71	3.55
	2	102.2868	108.6693	3.32	3.70
	3	100.0905	120.4925	3.19	3.51
	4	92.3152	87.9288	2.99	2.41
	5	88.4588	94.0023	3.05	3.28
Neural Network Only	6	92.4461	112.0605	3.22	3.44
	7	92.4647	116.6057	3.06	3.86
	8	90.351	131.677	2.99	4.00
	9	99.3026	135.0474	3.17	4.38
	10	98.2916	112.804	3.07	3.69
	11	100.2105	120.131	3.17	3.75
	12	92.0164	163.9895	3.02	4.34

As we can see here, in Table 5, the value of RMSE and percentage error is decreasing towards the increasing of number of neurons and the lowest RMSE and percentage error that could be achieve is at 4 neurons. The RMSE value of training phase is about 92.3152 GJ with percentage error of 2.99% and it being validate with 87.9288 GJ for RMSE and 2.41% error percentage. Afterwards the value of neurons increasing, an increasing of RMSE values and percentage error occurred.

To be concluded, the usage of 4 neurons is the best solutions to develop a neural network model. But by referring to the objectives of this research, the forecasting model must achieve the RMSE value less than 1% error. In order to achieve the objective, further step will be taken and the result will be discussed later in this chapter.

4.4 Repeat simulation using Neural Network with Genetic Algorithm

For this part, at first we only focus on unsorted data division. Then the result will be compare with neural network only result. The genetic algorithm part is being evaluated by first determination of number of generation. As depicted in Figure 4-3 the number of generation was increasingly improved the fitness along the run. To choose the best generation the highest generation that evokes the best fitness being selected. In this case, 500 generations is selected for experiment to be done because it offers the best result and there was no improvement in fitness observed after the period of 500 generations.
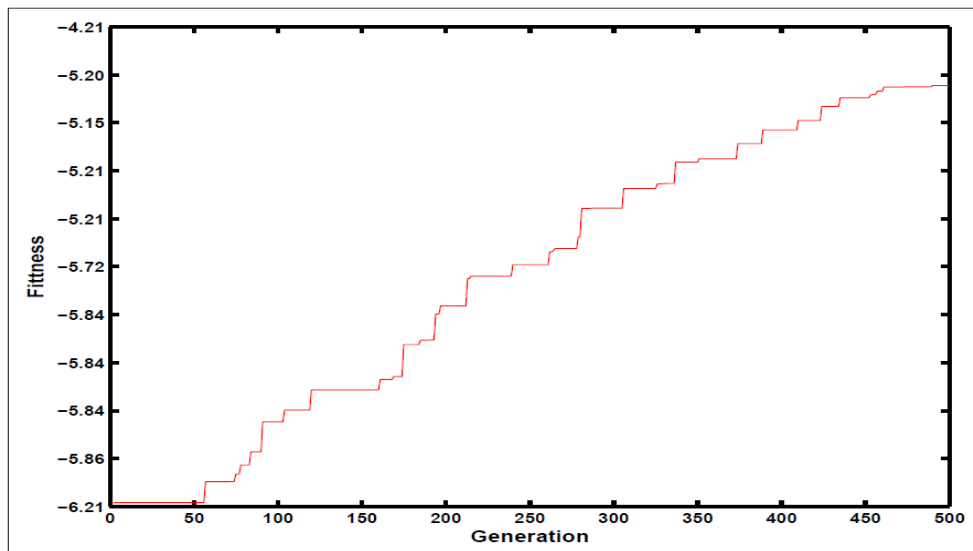

Figure 4-3: Performance of the Genetic Algorithm during the Run

After the determination of generations, simulation of unsorted data was being done. We can see from Table 6, the lowest RMSE and percentage error that this model can achieve is about 101.3498 GJ and 1.19% during training and 98.82 GJ and 1.53% during validation process. The result obtained is by the contribution of 7 neurons usage. To compare with result of neural network only model, we can see that genetic algorithm had improved the error function by minimize it to lowest it can be. Although it did not achieve the objective yet, but the percentage error produce is much more promising than neural network backpropogation model. The difference between NN with GA and NN only is about 1.8%.

Table 6: Unsorted Data Prediction using ANN with GA

Model and Data Division	No. of Neuron	RMSE (GJ)		Percentage of Error (%)	
		Training	Validation	Training	Validation
	1	132.1216	128.3349	4.42	2.74
	2	154.8168	105.7247	4.54	1.85
	3	207.9682	121.3889	5.67	2.01
Neural Network with Genetic Algorithm (Unsorted)	4	165.7613	116.4231	4.62	1.93
	5	128.6294	108.3074	2.70	1.83
	6	125.0922	108.2308	2.26	1.82
	7	101.3498	98.8200	1.19	1.53
	8	125.6143	103.0422	2.31	1.77
	9	142.4428	122.5053	4.44	2.02
	10	158.5719	139.7411	3.88	3.25
	11	247.4152	149.9982	5.90	3.83
	12	269.6133	146.6085	6.06	3.81

4.5   Analysis on Data Division

We have gone through the advantages of ANN combined with GA. The result obtained is much more reliable and efficient compare to the model of neural network only. In order to offer a better performance, this part is to test upon division of data to observed whether the split of data division influence the error percentage or not. Refer to methodology section; we had split the data into two parts: peak hour and non peak hour. For peak hour, the data is consisting of the energy consumption from 8 am to 5 pm. The result for prediction testing is according to the Table 4-3 below.

Table 7: Result upon Data Division (Peak Hour and Non Peak Hour)

Model and Data Division	No. of Neuron	RMSE (GJ)		Percentage of Error (%)	
		Training	Validation	Training	Validation
Neural Network with Genetic Algorithm (Peak Hour)	1	248.0941	229.5698	5.35	4.71
	2	155.5205	181.5512	2.97	3.37
	3	132.5228	106.7438	2.75	2.09
	4	135.9082	106.9345	2.94	2.13
	5	115.3105	103.0061	2.39	1.82
	6	105.0321	97.8154	1.36	1.09
	7	108.6980	96.5456	1.23	1.08
	8	145.6703	102.6915	3.27	1.77
	9	181.3570	103.3582	3.29	1.79
	10	193.2109	117.0231	3.82	2.96
	11	221.3172	118.3059	5.01	2.40
	12	231.8259	231.9879	5.26	4.16
Neural Network with Genetic Algorithm (Non Peak Hour)	1	136.9732	136.5880	4.74	4.43
	2	134.488	131.2936	4.02	3.35
	3	126.4133	108.8802	3.74	1.33
	4	117.0101	104.1099	2.31	1.29
	5	106.9063	106.2852	1.28	1.34
	6	106.5471	104.0899	1.25	1.24
	7	96.5471	107.2517	1.23	1.35
	8	103.9407	117.5329	1.16	1.46
	9	120.4090	123.0682	1.12	1.76
	10	153.2857	113.9302	1.96	1.93
	11	127.6582	115.3371	3.12	2.23
	12	184.3146	132.1372	4.38	3.73

Based on the result in Table 7, it is observed that for peak hour data set the RMSE values is decreasing with the increasing number of neurons then after it reached the optimum number of neurons, the RMSE values started increasing. Selection of neurons number depends on the least RMSE and percentage error values.

For the evaluation on the best value for the performance of the model, the RMSE value and the percentage of error are calculated. Based on the objective, the aim is to

develop an energy forecasting model that will produce accurate readings that gives minimum percentage of error of energy forecasting. Therefore, the main target is to design and develop a model that can give lowest possible error. In this case, the acceptable error between the predicted and actual value is 100 GJ. As we can see in Table 7, it is observed that all results of RMSE for all neurons tested are between 100 GJ to 200GJ for training phase and validation phase. For 6 and 7 neurons reading, it leads to the least RMSE for both training and validation for peak hour data set. The value for peak hour training RMSE at 7 neurons is 108.6980 GJ with 1.23% error and 96.5456 GJ for validation with 1.08% error. For non-peak hour readings, it shows that the lowest RMSE value for training phase is when 6 neurons being used which is 106.5471 GJ with 1.25% error and validation phase RMSE value is 104.0899GJ with 1.24% error.

In this Figure 4-4, comparison can be made clearly between neural network only and neural network with genetic algorithm. Validation data is the indicator to verify the training data whether neurons that being chose are suitable or not.
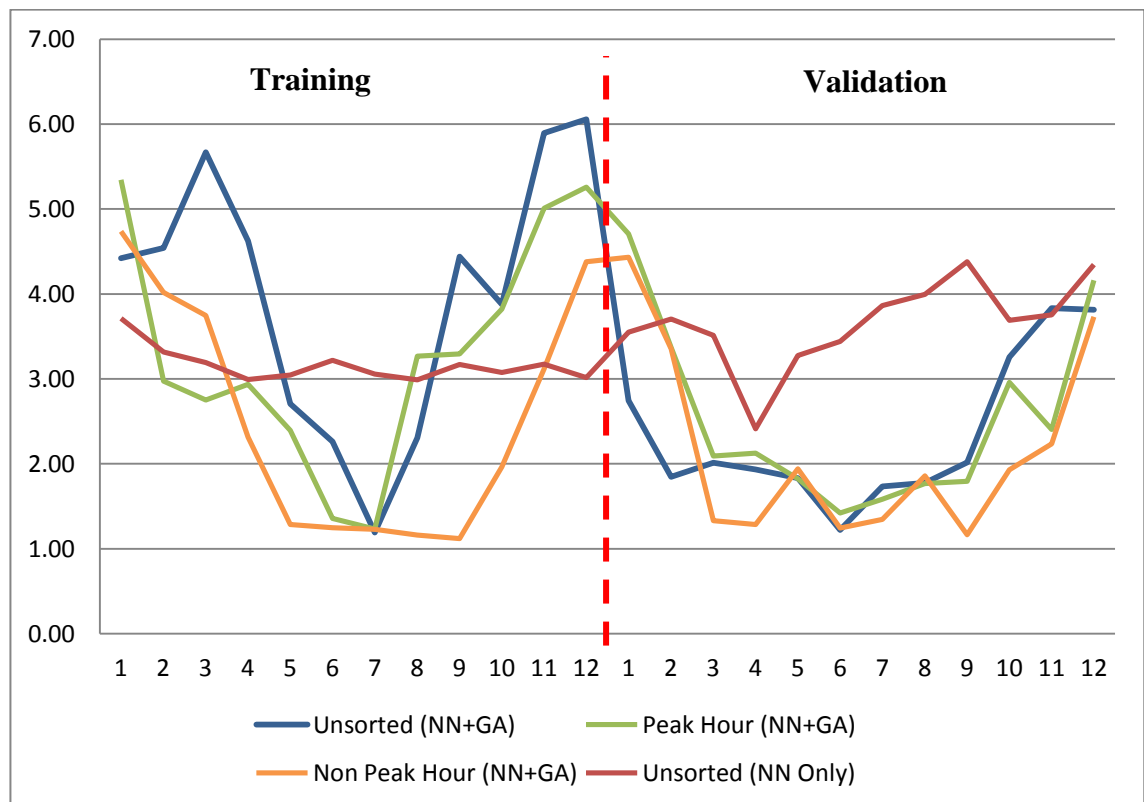
Figure 4-4: Summary of Prediction Model Performance

In a conclusion, the number of neurons that gives the lowest RMSE values and percentage of error is 6 and 7 neurons. Based on two sets of data division, the 7 neurons neural network model gives the least RMSE values and percentage of error for peak data but still considerable for non-peak hour . Still, although the RMSE values are less than 200 GJ but for error percentage is slightly bigger than aim which is below than 1% of error. For the process of literature review, assumption has been made that the readings of result is affected because of the limited data that can be collected because for neural network model, the number of data collected also one of contribution towards the performance. The more data can be collected, the less error of prediction model can be achieved.

Based on these investigations, the best parameters selected to build the neural network energy forecasting model are as shown in Table 8 below. The model is set up as shown in Figure 4.5.

Table 8: Parameters selected for neural network structure

Structure	Parameter
Network Architecture	Multilayer feedforward network
Learning Algorithm	Levenberg-Marquardt algorithm with Bayesian Regularization
No. of generation	500
Number of Neurons	7

Figure 4-5: Neural network energy with genetic algorithm forecasting model structure

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Chapter Overview

This chapter concludes the research findings and achievements based on the work completed for FYP 1 and FYP 2. Besides that, some recommendations for expansion and continuation of the project are included for improvement of the model developed.
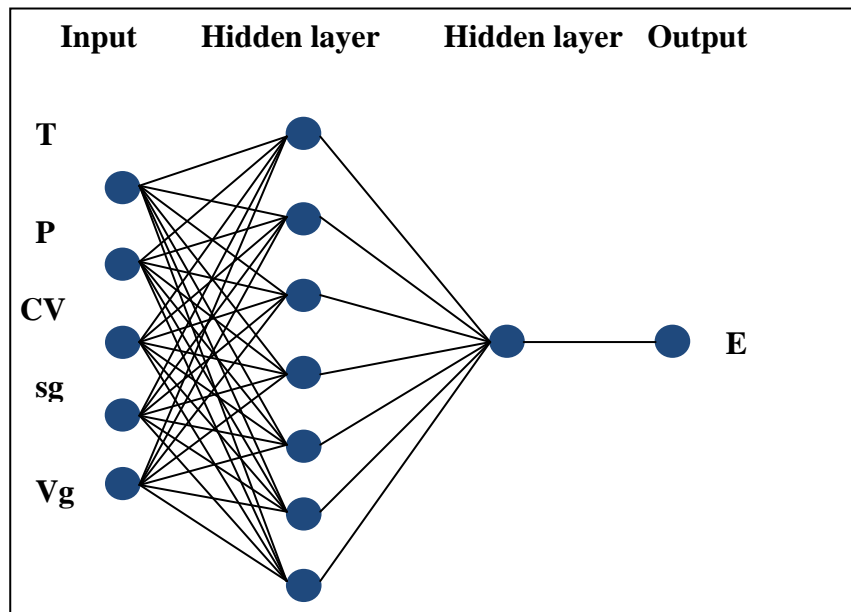
5.2   Conclusion

A neural network forecasting model of energy consumption has been successfully developed by constructing a multilayer feedforward network. The network is a three layer structure with one input layer, one hidden layer and one output layer. The learning algorithm used to train the network is the Levernberg-Marquardt with Bayesian Regularization algorithm with the percentage of 50% for training data and 50% for validation data. The combination of activation function is tan-sigmoid for hidden layer and pure-linear for output layer. The number of neurons needed in the hidden layer is 7 neurons. Besides that, the model is able to forecast energy with error less than 1 percent because of the filtration data according to peak hour and non peak hour condition.

As a conclusion, the objectives of this project can be achieved. A neural network with optimization of genetic algorithm forecasting model of energy consumption has been developed based on the process variables and it can be used as a verification tool of the outgoing gas flow. In the future, business credibility of PETRONAS and reliability of billing system can be maintained with the help of this prediction model. Furthermore, there is a possibility that it can be applied to other plants and the upstream sector in the oil and gas industry and has the potential of saving of millions of dollars to Malaysian oil and gas companies.

5.3   Recommendation

As what I can see, this project seems to have a brighter future to be developed and innovation should be done in order to achieve a successful forecasting model. For continuation of the project, another model should be study to compare the performance among models that can achieve the objectives recommend by PETRONAS. There are some other parameters that should be tested in order to prove the degree of performances. As an example, the parameter of epoch, mutation, generation, and selection should be the manipulated variable and make one of the variables being constant. The results of the model can be compared with the results of the existing studies of forecasting model and select the model that gives better performance.

On the other hand, it is also recommended to plan the phase of model development precisely and systematically. Proper planning of proceeding step can enhance the quality of result performance and procedure should be done as soon as possible because of the time constraints. Therefore, ones should brace themselves with as maximum knowledge as they can have before FYP 2 phase and can focus to forecasting model development during that phase.

REFERENCES

[1]   M. Abdella and T. Marwala, "The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database," *Computing and Informatics, 2005*, vol. 24, pp. 577-589, Oct. 2005.

[2]   P. D. Allison, "Multiple Imputation for Missing Data: A Cautionary Tale. In Sosiological Methods and Research.", vol. 28, pp. 301-309, 2000.

[3]   G. Gross and F. D. Galiana. "Short-term load forecasting", *Proceedings of the IEEE*, vol. 75, pp. 1558-1573, 1987.

[4]   H. Hahn, S. Meyer-Nieberg and S. Pickl. "Electric load forecasting methods: Tools for decision making", *European Journal of Operational Research*, pp. 902-907, 2009.

[5]   C. Chatfiled, "Time Series Forecasting. ", Chapman and Hall, 2000.

[6]   M. Jaeschke, "Thermodynamic research improves energy measurement of natural gas," Thermochimica Acta, vol. 382, no. 1-2, pp. 37-45, Jan. 2002.

[7]   P. J. Antsaklis, "Intelligent Learning Control", *IEEE Control System Magazine,* introduction to special issue on intelligcnt learning control, June 1995.

[8]   K. Y. Lee and J. H. Park, "Short-term Load Forecasting Using An Artificial Neural Network", *Transaction of Power System, 1992,* vol. 7, pp. 124-132, Feb. 1992.

[9]   A. Azadeh, S. F. Ghadrei and B. Pourvalikhan Nokhandan, "Short Term Load Forecasting by ANN", *IEEE Transactions*, pp. 1-5, 2009.

[10]   C. M. Zealand, D. H. Burn*, S. P. Simonovic, "Short term streamflow forecasting using artificial neural networks", *Journal of Hydrology*, pp. 32–48, Sept 1998.

[11]   J. Paris, "Neural Network Forecasting Model of Energy Consumption", *Final Year Project, Universiti Teknologi Petronas* pp. 1-68, September 2011.

[12]   J. Zaprianov and T. Atanasova, "Neural Network Integration in Control System Structures: A Case Study Towards Intelligent Control by Neural Network Integration", *Proceedings of the 12th IEEE,* pp. 167-172, July 1997.

[13]   R. Liaoa, H. Zhenga, Stanislaw Grzybowskib and Lijun Yanga, "Particle swarm optimization-least squares support vector regression based forecasting model on dissolved gases in oil-filled power transformers", *Electric Power Systems Research*, pp. 2074-2080, Aug. 2011.

[14]   Y. Li, T. Fang and E. Yu, "Short-term Electrical Load Forecasting Using Least Squares Support Vector Machines", pp. 230-234, 2002.

[15]   X. Zhao, G. Wang Gang, K. Zhao and D. Tan, "On-line least squares support vector machine algorithm in gas prediction", *Mining Science and Technology,* vol. 19, pp. 194-198, 2009.

[16]   T. Karna, F. Rossi and A. Lendasse, "LS-SVM Functional Network for Time Series Prediction", *ESANN'2006 proceedings,* pp. 1-6, 2006.

[17]   Jie Yang,L. P. Son and A. Bouzerdoum, "A training algorithm for sparse LS-SVM using compressive sampling", *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 2054-2057, 2010.

[18]   D. Basak, S. Pal and D. C. Patranabis, "Support Vector Regression", *Neural Information Processing – Letters and Reviews,* vol. 11, pp. 1-22, Oct 2007.

[19]   R. S. Sexton, R. E. Dorsey and J. D. Johnson, "Optimization of neural networks: A comparative analysis of the genetic algorithm and simulated annealing", *European Journal of Operational Research,* pp. 589-601, Feb. 1998.

[20]   P. F Pai and W. C Hong, "Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms", *Electric Power Systems Research,* pp. 417-425, April 2005.

[21]   V. Vapnik, S. Golowich and A. Smola, "Support vector machine for function approximation, regression estimation, and signal processing", *Advance Neural Information Process System,*vol. 9, pp. 281–287, 1996.

[22] A. J Smola, "Regression Estimation with Support Vector Learning Machines" *Master's thesis, Technische Universit" at M" unchen*, 1996

[23] J.A.K. Suykens, J. De Brabanter, L. Lukas and J. Vandewalle "Weighted Least Squares Support Vector Machines robustness and sparse approximation," *Neurocomputing special issue*, vol. 16, no. 3, pp. 512-534, 1999.

[24] M.J. Ismail and R. Ibrahim, "Selection of Network Architecture and Input Sensitivity Analysis for a Neural Network Energy Prediction Model," *2010 International Conference on Intelligent and Advanced Systems (ICIAS)*, pp. 1-6, Jun. 2010.

[25] P. S. Heckerlinga, B. S Gerbera, "Use of genetic algorithms for neural networks to predict community-acquired pneumonia", *Artificial Intelligence in Medicine Volume 30*, Issue 1, January 2004

[26] M. Oobaa, T. Hiranoa, J. Mogamia, R. Hiratab, Y. Fujinumab,"Comparisons of gap-filling methods for carbon flux dataset: A combination of a genetic algorithm and an artificial neural network", *Ecological Modelling*, Volume 198, Issues 3–4, 15 October 2006, Pages 473–486

[27] David J.C. Mackay, "Bayesian Interpolation," *Neural Computation*, vol. 4, pp. 415-447, 1992.

[28] Jan Poland, "On the Robustness of Update Strategies for the Bayesian Hyperparameter α," *Hokkaido University*, Japan, Nov. 2001.

APPENDICES

54

APPENDIX A

MATLAB CODING FOR NEURAL NETWORK AND GENETIC ALGORITHM

```matlab
%Clear workspace and command window
clc;

A = xlsread('unsortedV.xlsx');

%load data from workspace
x = A(:,1:5)';
y = A(:,6)';

%-------------------------------------------------------------------
---%
%prepocess the input and output [-1,1]
%-------------------------------------------------------------------
---%
[x_ti,x_s1] = mapminmax(x);    %INPUT training data
[y_ti,y_s1] = mapminmax(y);    %OUTPUT training data
% [x_v1,x_s2] = mapminmax(x_v);    %INPUT validation data
% [y_v1,y_s2] = mapminmax(y_v);    %OUTPUT validation data
%maximum and minimum value of TRAINING data
t = minmax(x_ti);

%-------------------------------------------------------------------
---%
%divide data into TRAINING and VALIDATION
%-------------------------------------------------------------------
---%
%get the number of input and number of data
numofdata=size(x,2); %number of data

% check if the number of data is even or odd value
if rem(numofdata,2)==0;
    numdata=numofdata-2;
else
    numdata=numofdata-3;
end

train_data = 0.5*numdata; %number of TRAINING data

validation_data =numdata-train_data; %number of VALIDATION data
numofvar = size(x,1); %number of input
numofout = size(y,1); %number of input

for m=1:numofvar
    for n=1:train_data
    x_t(m,n)=x(m,n);
    end
end

for m=1:numofvar
    for n=1:validation_data
```

```matlab
x_v(m,n)=x(m,n+train_data);
    end
end

for m=1:numofout
    for n=1:train_data
    y_t(m,n)=y(m,n);
    end
end

for m=1:numofout
    for n=1:validation_data
    y_v(m,n)=y(m,n+train_data);
    end
end

%-------------------------------------------------------------------
---%
%prepocess the input and output [-1,1]
%-------------------------------------------------------------------
---%
x_t1 = mapminmax('apply',x_t,x_s1);   %prepare INPUT data for
training
y_t1 = mapminmax('apply',y_t,y_s1);   %prepare OUTPUT data for
training
x_v1 = mapminmax('apply',x_v,x_s1);   %prepare INPUT data for
validation
y_v1 = mapminmax('apply',y_v,y_s1);   %prepare OUTPUT data for
validation
%maximum and minimum value of TRAINING data
t1 = minmax(x_t1);


%-------------------------------------------------------------------
---%
%set network properties
%-------------------------------------------------------------------
---%
%number of neurons for layer 1 and layer 2
for neuron_1 = 1 : 12;

neuron_1 = 1; %number of neurons for layer 1
neuron_2 = 1; %number of neurons for layer 2

%network and parameters
net=newelm(x_i,y_i,neuron_1,{'tansig','tansig'},'traingdx');
net.trainParam.show = 50;
net.trainParam.lr = 0.1;
net.trainParam.epochs = 1000;
net.trainParam.goal = 0.001;
net=init(net);

%checking the weights and biases (make sure all are 0)
net.IW{1,1}; %weights of 1st layer
net.LW{2,1}; %weights of 2nd layer
net.b{1}; %bias of 1st layer
net.b{2}; %bias of 2nd layer
```

```matlab
%----------------------------------------------------------------%
%train the network
%----------------------------------------------------------------%
[net,tr]=train(net,x_t1,y_t1);

%[net,y,e]=adapt(net,x_t1,y_t1);
%net.adaptParam.passes=100;




%----------------------------------------------------------------%
%simulate the network
%----------------------------------------------------------------%
%simulate the network with TRAINING data
xtest_t = mapminmax('apply',x_t,x_s1);  %prepare input data for
training
ytrain = sim(net,xtest_t);  %simulate the network
ytrain1 = mapminmax('reverse',ytrain,y_s1); %descale the output
%calculate the different between the actual and predicted
temperature value
etrain=y_t-ytrain1;
%simulate the network with VALIDATION data
xtest_v = mapminmax('apply', x_v, x_s1);    %prepare input data
for training
yvalid=sim(net,xtest_v); %simulate the network
yvalid1 = mapminmax('reverse',yvalid,y_s1);%descale the output
%calculate the different between the actual and predicted
temperature value
evalid=y_v-yvalid1;


%----------------------------------------------------------------%
% Genetic Algorithm part
%----------------------------------------------------------------%
% for the VALIDATION data

evalidga= evalid;
yvalid1 = validationga(evalidga);
[evalidgamax f]=ga(@validationga,2);
%y_v=evalidga(1);

%new evalidga
new_evalidga = -(evalidga- evalidgamax(:,2));
new_evalidga = new_evalidga(:,1)';
new_yvalid1 = y_v-new_evalidga;



% for the TRAINING data



etrainga= etrain;
ytrain1 = trainingga(etrainga);
[etraingamax f]=ga(@trainingga,2);
%y_t=etrainga(1);

%new etrainga
new_etrainga = -(etrainga- etraingamax(:,2));
new_etrainga = new_etrainga(:,1)';
```

```matlab
%-------------------------------------------------------------------
-----%
%error analysis
%-------------------------------------------------------------------
-----%
%error analysis for the VALIDATION data

rmse_valid (neuron_1)= sqrt(mse(evalid)); %mean square error
percenterror_valid (neuron_1)= mean((abs(y_v-
yvalid1)./y_v).*100); %percent error



%error analysis for the TRAINING data
rmse_train(neuron_1) = sqrt(mse(etrain)); %mean square error
percenterror_train(neuron_1) = mean((abs(y_t-
ytrain1)./y_t).*100); %percent error

%error analysis for the VALIDATION data using GA

rmse_validga (neuron_1)= sqrt(mse(new_evalidga)); %mean square
error
percenterror_validga (neuron_1)= mean(abs(y_v-
new_yvalid1)./y_v).*100; %percent error

%error analysis for the TRAINING data using GA
rmse_trainga(neuron_1) = sqrt(mse(new_etrainga)); %mean square
error
percenterror_trainga (neuron_1)= mean(abs(y_t-
new_ytrain1)./y_t).*100;


end


%-------------------------------------------------------------------
-----%
%plot graph
%-------------------------------------------------------------------
-----%
% plot the actual and predicted energy from VALIDATION data
    subplot(2,2,1);
    plot (new_yvalid1,'r');
    hold on;
    plot (y_v,'b');
    xlabel('No of Data');
    ylabel('Energy (GJ)');
    title('Output of NN+GA model for Energy (Validation Data)');
    legend('Predicted Energy','Actual Energy');
    grid on;

% %plot the different between the actual and predicted energy
from VALIDATION data
    subplot(2,2,2);
    plot(new_evalidga,'*');
    xlabel('No of data');
    ylabel('Error (GJ)');
    title('Error between Actual Energy and Predicted Energy
(Validation Data)');
    grid on;
```

```matlab
% %plot the actual and predicted energy from TRAINING data
    subplot(2,2,3);
    plot (new_ytrain1,'r');
    hold on;
    plot (y_t,'b');
    xlabel('No of Data');
    ylabel('Energy (GJ)');
    title('Output of NN model for Energy(Training Data)');
    legend('Predicted Energy','Actual Energy');
    grid on;
%
% %plot the different between the actual and predicted energy from
TRAINING data
    subplot(2,2,4);
    plot(new_etrainga,'*');
    xlabel('No of Data');
    ylabel('Error (GJ)');
    title('Error between Actual Energy and Predicted Energy
(Training Data)');
    grid on;


% plot the regression between the actual and predicted energy from
VALIDATION
% data

    plotregression(yvalid1,y_v)
    hold on;
    ylabel('Energy (GJ)');
    title('Output Regression of NN model for Energy (Validation
Data)');
    legend('Predicted Energy','Actual Energy');

    plot the regression between the actual and predicted energy
from TRAINING
    data

    plotregression(ytrain1,y_t)
    hold on;
    ylabel('Energy (GJ)');
    title('Output Regression of NN model for Energy (Training
Data)');
    legend('Predicted Energy','Actual Energy');
%
```