PREDICTIVE MODEL AND NEAR INFRARED SPECTROSCOPY IN PREDICTING
THE DIESEL FUEL PROPERTIES

HASAN ALI GAMAL AL-KAF

A project report submitted in partial
fulfillment of the requirement for the award of the
Degree of Master of Electrical Engineering

Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia

JANUARY 2018
Dedicate this work to my beloved mother and father and my brothers and my close friend Ahmed
ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to ALLAH Almighty for His blessing to this great achievement. Without His grace and comparison, none of his would have been possible.

I would like to express my deepest appreciation to my supervisor, Dr Chia Kim Seng for the guidance, enthusiasm for the entire progress of this project.

Special thanks to my family members and friends who have been helpful by offering comment and advice ensuring the success of this research.
ABSTRACT

Monitoring the diesel fuel properties play an important role in the performance of vehicle engines. Near-infrared (NIR) technology has been investigated as an alternative to monitor the diesel fuel properties. NIR spectroscopy shows an enormous potential for quantitative analysis of complex samples by coupling with artificial neural networks (ANNs). Although a single layer ANN shows promising in the establishing better relationship between a component of interest and NIR spectrum, a different algorithm for updating weight that has been proved to improve the performance of the multilayer could further reveal the potential of single linear layer ANN in NIR spectroscopic analysis. Therefore, this study investigates the performance of a single layer ANN that trained with Levenberg-Marquardt (SLM) and that trained with Scaled Conjugate Gradient (SSCG) and compares the proposed methods with multilayer ANN that trained with same learning algorithms. Results were evaluated and discussed with previous studies that used the same data sets to establish the relationship between the NIR spectral data and diesel fuel properties. Finding depicts that the proposed SLM and SSCG were capable of predicting the diesel fuel properties using NIR spectrum without data reduction, and achieving better accuracy in predicting the diesel fuel properties compared with other recent methods. In addition, using a proposed genetic algorithm for data reduction to improve the predictive model of the proposed method.
ABSTRAK

Pemantauan sifat bahan api diesel memainkan peranan penting dalam prestasi enjin kenderaan. Kajian telah menunjukkan teknologi infra-merah dekat (NIR) boleh digunakan sebagai kaedah alternatif untuk memantau sifat-sifat bahan api diesel. Spektroskopi NIR menunjukkan potensi besar untuk menganalisis kuantitatif sampel yang kompleks dengan gandingan rangkaian saraf tiruan (ANN). Walaupun satu lapisan ANN menjanjikan hubungan yang lebih baik antara komponen berkaitan dan spektrum NIR, algoritma yang berbeza untuk mengemaskini berat telah membuktikan dapat meningkatkan prestasi pelbagai lapisan ANN dapat menunjukkan potensi lapisan linear tunggal ANN dalam analisis spektroskopi NIR. Oleh itu, kajian ini dilakukan untuk menyiasat prestasi lapisan tunggal ANN yang dilatih dengan Levenberg-Marquardt (SLM) dan Gredent Conjugate Scaled (SSCG) dan kemudian dibandingkan dengan kaedah pelbagai lapis ANN yang dilatih dengan algoritma yang sama. Hasil keputusan telah dinilai dan dibincangkan dengan kajian sebelumnya yang menggunakan set data yang sama untuk mewujudkan hubungan antara data spektrum NIR dan bahan api diesel. Penemuan kajian menunjukkan bahawa SLM dan SSCG yang dicadangkan mampu meramalkan sifat bahan api diesel menggunakan spektrum NIR tanpa pengurangan data, dan mencapai ketepatan yang lebih baik dalam meramalkan sifat bahan api diesel berbanding dengan kaedah sebelum ini. Di samping itu, penggunaan algoritma genetik untuk pengurangan data dapat memperbaiki ramalan kaedah menggunakan model yang dicadangkan.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>i</td>
</tr>
<tr>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td>DICATION</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
</tr>
<tr>
<td>CONTENTS</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xiii</td>
</tr>
<tr>
<td><strong>CHAPTER 1</strong> INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Project Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Objectives</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Scope</td>
<td>5</td>
</tr>
<tr>
<td><strong>CHAPTER 2</strong> LITERATURE REVIEW</td>
<td>6</td>
</tr>
</tbody>
</table>
2.1 Introduction 6
2.2 Important of diesel fuel properties 6
2.3 Near Infrared Spectroscopy (NIRS) 7
2.4 Multivariable Calibration Analysis 8
  2.4.1 Classical Least Squares 9
  2.4.2 Inverse Least Square 10
  2.4.3 Partial least squares (PLS) 10
  2.4.4 Multiple linear regression (MLR) 11
  2.4.5 Artificial neural network 12
2.5 Genetic Algorithm 13
  2.5.1 Important consideration 14
  2.5.2 Genetic algorithm and near infrared spectroscopy 16
  2.5.3 Previous Research on Genetic Algorithm with other Regression 16
  2.5.4 Neural network coupled with Genetic algorithm 19
2.6 Summary 20

CHAPTER 3 METHODOLOGY 21
3.1 Introduction 21
3.2 Flow of the works 21
3.3 Theory and algorithm 23
  3.3.1 Single layer network 23
  3.3.2 Multilayer network 24
  3.3.3 Levenberg-Marquardt (LM) 25
  3.3.4 Scaled Conjugate Gradient (SCG) 25
  3.3.4 Single Layer Neural Network 26
3.4 Experimental
   3.4.1 Efficient of dataset 27
   3.4.2 Parameter setting for single layer and multilayer 28
3.5 Genetic algorithm for feature selection 29
3.6 Evaluation the performance 32
   3.6.1 Root mean square error 32
   3.6.2 Correlation Coefficient 33
CHAPTER 4 RESULTS AND DISCUSSION 34
4.1 Introduction 34
4.2 Optimization of iteration of SLM/SCG 34
4.3 Evaluation of the single layer and multilayer trained with lm and SCG 37
4.4 Evaluation of the single layer with previous studies 41
4.5 Genetic algorithm with single layer 43
   4.5.1 First Genetic algorithm with single layer 43
       trained with SCG 43
   4.5.2 First Genetic algorithm with single layer 45
       trained with LM 45
   4.5.3 Second Genetic algorithm with single layer 45
       trained with SCG 45
   4.5.4 Second Genetic algorithm with single layer 45
       trained with LM 45
CHAPTER 5 CONCLUSION AND RECOMMENDATIONS 47
5.1 Conclusion 47
5.2 Recommendation 48
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>TABLE</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Previous Research on Genetic Algorithm with other Regression</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Selection method, Generation and fitness function of previous research</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of preparation of chromosomes</td>
<td>29</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison between single layer and multilayer</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison between Adaline and proposed methods</td>
<td>41</td>
</tr>
<tr>
<td>4.3</td>
<td>Compare the performance of the proposed method and Partial Least Square (PLS) extreme learning machine (ELM) and boosting extreme learning machine (Booting ELM)</td>
<td>42</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison between proposed method and genetic inverse least squares</td>
<td>44</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Objective of Multivariable Regression</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Type of Multivariate Calibration Models</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>Producer of Genetic Algorithm</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Overall works of the methodology</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>General Description of SLM and SSCG models</td>
<td>27</td>
</tr>
<tr>
<td>3.3</td>
<td>Flowchart of steps of Genetic algorithm</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>The performance of the single layer neural network for different properties with different iteration for LM algorithm to find optimized iteration</td>
<td>35</td>
</tr>
<tr>
<td>4.2</td>
<td>The performance of the single layer neural network for different properties with different iteration for SCG algorithm to find optimized iteration</td>
<td>36</td>
</tr>
<tr>
<td>4.3</td>
<td>The regression plot of the proposed single layer ANN trained by LM for prediction</td>
<td>37</td>
</tr>
<tr>
<td>4.4</td>
<td>The regression plot of the proposed single layer ANN trained by SCG for prediction</td>
<td>38</td>
</tr>
<tr>
<td>4.5</td>
<td>The performance of the multi-layer ANN that trained by SCG and LM algorithms in predicting the boiling point</td>
<td>39</td>
</tr>
<tr>
<td>4.6</td>
<td>The performance of first and second genetic algorithm for different LM and SCG algorithm</td>
<td>46</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Project Gantt chart</td>
<td>56</td>
</tr>
<tr>
<td>B</td>
<td>First Genetic Algorithm with Single Layer</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Trained with SCG</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>First Genetic Algorithm with Single Layer</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Trained with LM</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Second Genetic Algorithm with Single Layer</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Trained with SCG</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Second Genetic Algorithm with Single Layer</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Trained with LM</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Project Background

Near-Infrared (NIR) spectroscopy has been widely proposed as an alternative for the determination of varies chemical compounds, e.g. the products of petroleum refining and petrochemicals, food products, and pharmaceutical [1–3]. It has proved its efficiency for laboratory and industrial applications with the advantages of rapid, low cost, and non-invasive. One of the promising usages of NIR spectroscopy is in petroleum industry [4]. This is because petroleum refining and petrochemicals consist of hydrocarbons, which can be quantified by NIR sensing technique. Classical calibration methods i.e. linear and multiple regression is important in helping researchers or users to understand the relationship between the dependent and independent variables. Unfortunately, these classical calibration approaches are unable to directly model complex and high dimensional NIR data. Thus, pre-processing and data reduction strategies are necessary to be performed prior the use of classical calibration approaches so that sophisticated NIR spectral data can be modeled.

NIR spectroscopy shows an enormous potential for quantitative analysis of complex samples by coupling with artificial neural networks (ANNs). Many studies use the NIR spectroscopy coupled with multilayer neural network (MNN) for quantifying the concentrations of urea, creatinine, glucose, oxyhemoglobin [5], the rice wine age [6], the soluble solid content of intact pineapple [7], the fiber contents of textile mixture [8], and the cephalexin [9]. The applicability of the ANN approach has been increased due to the
advances of learning algorithms such as Levenberg-Marquardt, Scaled Conjugate Gradient, Gradient Descent, and One Step Scant.

Nevertheless, multilayer neural network (MNN) is complex in terms of determining its elements i.e. the number of a hidden neuron, the initial weight, and other learning parameters. In most cases, the idea about the right number of hidden neurons might not be clear to identify. Therefore, this leads to many trial-and-error times for finding a right model for an application [10, 11]. Moreover, an excessive number of hidden layers or hidden neuron could negatively affect the generalization of MNN that leads to over-fitting issues [12].

Recently, boosting extreme learning machine (ELM) for single hidden layer feedforward neural networks has been introduced to the local minimum without complex parameters tuning [13]. The boosting ELM has two parameters to be optimized i.e. the activation function and the number of hidden nodes. The inherent characteristics of simple structure, excellent predictive performance, and high learning speed attribute the superiority of boosting ELM [13]. Both weights and biases for hidden nodes are randomly generated without iteratively adjusting in ELM1 [4, 15]. Even though this will increase the learning speed and reduce the number of parameters that need to be optimized, the initialization of the hidden layer biases and input weights that done randomly may make ELM unstable in practices and the complex structure of ELM [15, 16]. Additionally, the complexity of ensemble modeling strategy e.g. the boosting ELM that consists of many sub-models could be a barrier for researchers to understand the fundamental relationship between near infrared wavelengths and the component of interest.

The simplest structure of ANN i.e. single layer ANN that coupled with Levenberg-Marquardt [17] or Gradient Descent technique [18] has shown a potential to achieve faster convergence with lower computational complexity implemented in system identification e.g. real-time adaptive control and online system identification. On the other hand, a single linear layer that trained with Widrow-Hoff delta rule has been successfully implemented to model high dimensional NIR spectral data to predict the boiling point of diesel fuel without any data reduction approach, and achieved better performance compared with principal component regression (PCR) and Partial least square (PLS)[19]. However, the single layer that trained with Widrow-Hoff delta rule depends on two important factors
i.e. the adaption cycle and learning rate. Consequently, a trial and error approach is needed to optimize these two parameters to avoid overfitting problems. Perhaps different algorithms e.g. Levenberg-Marquardt, Scaled Conjugate Gradient for updating weight that has been proved to improve the performance of the multilayer could further reveal the potential of single linear layer ANN in NIR spectroscopic analysis.

In summary, the limitation of the current approaches i.e. the updating weight is a crucial task to improve the accuracy of a single layer ANN. Second, for our best of knowledge different algorithms such as LM and SCG has not yet applied for the single layer ANN in NIR studies. Third, the number of hidden neurons of both MNN and ELM are required to be optimized while single layer ANN e.g. Adaline does not need to optimize this parameter. Fourth, the predictive accuracy is not proportional to the model complexity. This is because a multilayer ANN that gives a good training accuracy may lead to worse generalization than a single layer in some cases. In addition, Spectral data of near infrared consisting of hundreds and even thousands of absorbance values per spectrum. Some of these values are irreverent and not interest and existing of outlying samples and nonlinearity characteristics of NIR spectrum can degrade the performance of a predictive model significantly. Genetic algorithm widely used as features selection to improve the predictive models.

1.2 Problem Statement

Monitoring the diesel fuel properties play an important role in the performance of vehicle engines. Near infrared (NIR) technology has been investigated as an alternative to monitoring the diesel fuel properties. Spectral data of Near infrared consisting of hundreds and even thousands of absorbance values per spectrum. Some of these values are irreverent and not interest and existing of outlying samples and nonlinearity characteristics of NIR spectrum can degrade the performance of a predictive model significantly. Therefore, NIR spectroscopy coupling with artificial neural networks (ANNs) has improved the performance of predicted models. The most widely used is multilayer neural network which shows a good performance but multilayer neural network (MNN) is not a simple model in terms of determining its elements i.e. the number of a hidden neuron, the
initial weight, and other learning parameters. In most cases, the idea about the right number of hidden neurons might not be clear to identify. Therefore, this leads to many trial-and-error times for finding a right model for an application. Moreover, an excessive number of hidden layers or hidden neuron could negatively affect the generalization of MNN that leads to over-fitting issues.

Recently, extreme learning machine (ELM) for single-hidden layer feedforward neural networks (SLFNs) has been introduced to the local minimum without complex parameters tuning [19]. The ELM has two parameters to be optimized i.e. the activation function and the number of hidden nodes. The inherent characteristics of simple structure, excellent predictive performance, and high learning speed attribute the superiority of ELM [19]. Both weights and biases for hidden nodes are randomly generated without iteratively adjusting in ELM. Even though this will increase the learning speed and reduce the number of parameters that need to be optimized, the initialization of the hidden layer biases and input weights that done randomly may make ELM unstable in practices.

Single layer trained with withdraw Hoff (Adaline) is simple structure and has a strong relationship between the near infrared spectrum and the physical properties, but the performance does not give a good performance compared with MNN. In addition, Adaline depends on two important factors i.e. the adaption cycle and learning rate. Consequently, a trial and error approach is needed to optimize these two parameters to avoid overfitting problems. Perhaps different algorithm for updating weight that has been proved to improve the performance of the multilayer could further reveal the potential of single linear layer ANN in NIR spectroscopic analysis.

1.3 Objectives

In this thesis, the overall goal is to predict boiling point of diesel using near infrared spectral data and predictive model. In order to achieve this, the following objectives are listed:

1. Develop a single layer ANN and genetic algorithm based ANN in NIR spectroscopic analysis.
2 To establish the relationship between the NIR spectrum and the diesel fuel properties using the developed ANN.
3 To evaluate the performance of the proposed single layer ANN with multilayer networks that trained with same learning algorithms without data reduction, and the previous studies that used same NIR data.

1.4 Scope

In this thesis, the overall scopes are to provide a clear idea on how genetic algorithm and single layer trained with Levenberg-Marquardt, Scaled Conjugate Gradient is a good technique for optimization the diesel fuel properties by using near infrared spectroscopy. In order to achieve this, the following scopes are listed below:

1 The NIR spectral data of diesel fuel samples that measured at the Southwest Research Institute (SWRI) will use in this study.
2 Data processing and modeling will use by MATLAB (version R2015b, win64)
3 A single layer trained with Levenberg-Marquardt, Scaled Conjugate Gradient will use as a multivariable technique to predict diesel fuel properties based on near infrared spectrum.
4 Multilayer neural network (MNN) is used as regression to compare the accuracy with proposed methods.
5 Evaluated and compared with previous studies that used same NIR datasets such as partial last square (PLS), adaptive linear neuron(Adaline), extreme learning machine (ELM), Boosting(ELM) and inverse genetic inverse least squares
6 A genetic algorithm is used as feature selection for NIR Spectral data of diesel samples.
CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, the general concept that is related to predicting model using near infrared spectroscopy was reviewed and the significance that motivates in conducting this research was presented in the subsequent chapters of this thesis. First, Section 2.2 depicts the importance of diesel fuel properties. Second, Section 2.3 a brief about near infrared spectroscopy as well as the usage of near infrared spectroscopy in diesel fuel were discussed and presented. Then, Section 2.4 Multivariable techniques were discussed and Section 2.5 focusing on the artificial neural network. Finally, Section 2.6 The Genetic technique include the producer of the genetic algorithm, an important consideration should be taken while using a genetic algorithm and finally the previous research that combines genetic algorithm with other regression were discussed and focusing on an artificial neural network coupled with a genetic algorithm.

2.2 Important of Diesel Fuel Properties

The significance of the different fuel parameters and the ASTM reference method used to determine them are described below. The boiling point, the appearance of components of a high boiling point in fuels can affect the degree of formation of solid combustion products. The freezing temperature is a significant indicator to emphasize
the smooth supply of fuel in an engine and coming to the flowing properties of diesel fuel, the viscosity and density affect these properties in the pipeline. One of the most fundamental physical properties is density which, in conjunction with other properties, can be used to characterize both heavy and light fractions of petroleum and petroleum products. Also, determining the density (or relative density of petroleum and its products) is important for the process to convert the measured volumes to the volumes at the standard temperature of 15 °C. Cetane number is another important property, fuel can be divided into different brands based on the cetane number. In compression ignition engines, the cetane number also, provides a measure of the ignition characteristics of diesel fuel oil, and is, also, used by petroleum refiners, engine manufacturers, marketers and in commerce as a primary specification measurement related to matching fuels with engines. The sediment in the engine is increased by the aromatics, and toxic substances are produced by aromatic hydrocarbons.

2.3 Near Infrared Spectroscopy (NIRS)

Near-Infrared (NIR) spectroscopy has been widely proposed as an alternative for the determination of various chemical compounds, e.g. the products of petroleum refining and petrochemicals, food products, and pharmaceutical [1–3]. It has proved its efficiency for laboratory and industrial applications with the advantages of rapid, low cost, and non-invasive. One of the promising usages of NIR spectroscopy is in petroleum industry [4]. This is because petroleum refining and petrochemicals consist of hydrocarbons, which can be quantified by NIR sensing technique. Near infrared reflectance spectra (NIRS) has become an effective method for rapid and real-time analysis of fuel properties [20-30]. Organic functional groups that contain hydrogen can form broad absorption bands within near infrared wavelength range. Diesel fuel mainly consists of organic materials such as cetane, which contains a large number of functional groups. Therefore, it is appropriate to apply NIRS to rapid analysis of diesel fuel properties.
Considerable progress has been reported over the last decade [21-29] e.g., Felicio measured the composition of fuel by NIRS [27]; Coope predicted density, viscosity, and boiling point at 50% recovery of diesel fuel using NIRS [28]; Dehui Wu tried online measurement for diesel fuel with NIRS [29].

2.4 Multivariable Calibration Analysis

Multivariate Calibration analysis is very important of NIR data because of the overlapping nature of NIR spectra for condensed phase and the sensitivity of these spectra. These methods are used to achieve robustness and to enhance selectivity by basing measurements on an analysis of the full spectrum. The main objective of the multivariate calibration method is to relate a chemical or physical property of interest to the spectral information encoded across multiple wavelengths by establishes a mathematical model. There are two-step procedures. The first step is to collect a set of standard sample and this sample is pre-determined by an established independent reference assay. Second is to treat the sample spectra and reference as the calibration data for the purpose of establishing a mathematical can correlate the target property to feature in the spectral data set. Finally, the mathematical model can use to predict the model property for subsequent unknown samples. Figure 2.1 shows the producer and the objective of Multivariate Calibration analysis.

![Figure 2.1: Objective of Multivariable Regression](image-url)
Figure 2.2 shows a variety of multivariate calibration methods are available for the analysis of NIR spectral data. Classical least-squares (CLS) and inverse least-squares (ILS) regressions, multiple linear regression (MLR), principal component analysis and regression (PCA and PCR), partial least squares (PLS) and net analyte signal (NAS).

![Diagram of Multivariate Calibration Models]

**Figure 2.2: Type of Multivariate Calibration Models**

### 2.4.1 Classical Least Squares

Classical Least Squares (CLS) models the spectral variances based on the Beer-Lamber Law. (CLS) and inverse least-squares (ILS) regressions which are often referred as K and P matrix methods, respectively [31-32]. The constituent’s concentration and the corresponding are an important parameter as shown in Equation 2.1 [32-33]. The constituent spectra $S$ can be estimated simply with an ordinary least squares pseudo-inversion of $C$ as shown in Equation 2.2. To predict the concentration of analytes in a new sample from collected spectra is shown in Equation 2.3 and the pseudo-inversion of estimated constituent spectra $S$ is defined in equation.

$$ R = CS^T + E $$  \hspace{1cm} (2.1)

$$ S^T = C^+ R $$  \hspace{1cm} (2.2)

$$ C = S^+ R_{Unknown} $$  \hspace{1cm} (2.3)

Where $R$ is the vector of sample spectra, $S$ contains the spectra of each individual constituent in $R$ and is determined by regression, $C$ is the corresponding concentration matrix of analyses of interest and $E$ is the error matrix associated with the calibration.
CLS manipulate the spectra and concentration in an easy way and the constituent spectra estimate through the regression. The disadvantage of this method is the need to have a component spectrum for each element which composites the spectrum. A new approach called advanced CLS (ACLS) could reach the robust modeling within a complex matrix.

2.4.2 Inverse Least Square

The concentration of sample predicted quantitatively form sample spectra [34]. The equation 2.4 shows the basic model for ILS. The vector P is designed to selective to the analyte for modeling and orthogonal to all other components of spectral variation in the spectra matrix R. the regression vector is found in equation 2.5. The advantage of this method is that has a purely mathematical construction.

\[ C = RP + E \]  \hspace{1cm} (2.4)
\[ P = R^+C \]  \hspace{1cm} (2.5)

Where R is the matrix of spectral data, \( R^+ \) is the pseudo-inverse of the spectral matrix R. The vector P is the correlation between the spectral matrix R and the concentration matrix C. E is the matrix of concentration errors that are not included by the model.

2.4.3 Partial Least Squares (PLS)

The Partial least squares were proposed by H.Wold in the 1960s [35]. PLS can estimate the model between spectral variance and concentration in linear regression [36-37]. The basic concept of this method is to decompose the spectral matrix R and the concentration matrix C into loading and score. The second step is to establish the correlation of covariance as shown in equation 2.8. To ensure maximum correlation with the loadings from the concentration, the latent variables from spectral matrix R in PLS are developed as a linear combination of the spectral information.
The advantage of this is to enhance a subspace in the spectral matrix $R$ that is better to model concentration. The relevant equations are:

$$R = TP + E$$
$$C = UQ + F$$

Where $T$ and $U$ are score matrices of $R$ and $C$ and, $P$ and $R$ are the respective loading matrices. $E$ and $F$ are the errors associated with the corresponding $R$ and $C$ matrices. The normalized first weight vector which about the connection is between matrices $R$ and $C$ is constructed in order to maximize correlation with the concentration matrix $C$. The $w$ is computed as follow.

$$W_1 = \frac{R^T C}{\|R^T C\|}$$

(2.8)

Consequently, the regression vector related to the first factor can be estimated as

$$W_1 = \frac{C^T (R^T W_1)}{\|(R W_1)^T R W_1\|}$$

(2.9)

Subsequently, the algorithm repeats for additional factors by calculating $w_2$ and further regression vectors with the concentration and spectra residuals based on the first factor. For concentration predictions, the calibration vector matrix is used according to Equation 2.10.

$$C^\wedge = XB$$

(2.10)

Because the PLS model extracts the latent variables by building covariance between the spectral and concentration matrices, the PLS method efficiently minimizes the impact of interferences. For this reason, PLS methods represent popular tools for multivariate measurements in complex samples.

### 2.4.4 Multiple Linear Regression (MLR)

This technique used an ordinary least square algorithm to compute the pseudo-inverse of spectra as shown in equation 2.11. The method models a linear relationship between one dependent variable or more [38].

$$R^+ = (R^T R)^{-1} R^T$$

(2.11)

The pseudo-inverse utilized the inversion of RTR, This technique suffers from its instinctive shortcomings and is very sensitive to the conditioning of the spectral matrix.
Estimation of the true inversion of RTR is difficult and inaccurate because of the collinearity existing in the spectral matrix R. Spectral feature are highly overlapped between components and due to this reason, the resulting collinearity in the spectral matrix R diminishes performance of the MLR approach. the number of spectral variables is more than the independent spectra in the matrix R which result in inaccurate estimation of the regression vector. In order to compensate the limitations, a latent variable algorithm such as PCA and PLS have been developed.

2.4.5 Artificial Neural Network

NIR spectroscopy shows an enormous potential for quantitative analysis of complex samples by coupling with artificial neural networks (ANNs). Many studies use the NIR spectroscopy coupled with multilayer neural network (MNN). The applicability of the ANN approach has been increased due to the new advances and recent popularization of learning algorithms such as Levenberg-Marquardt (LM), Scaled Conjugate Gradient (SCG), Gradient Descent (GD), and One Step Scant (OSS). For example, ANN has been investigated and implemented in predicting the pre-oxidation efficiency of refractory gold concentrate [39] the blood glucose [40]. The emissions from swine buildings [41], the software defect [42], the risks of capital flow [43], and the sleep disorders [44]. Nevertheless, Multilayer neural network (MNN) is not a simple model in terms of determining its elements i.e. the number of a hidden neuron, the initial weight, and other learning parameters. In most cases, the idea about the right number of hidden neurons might not be clear to identify. Therefore, this leads to many trial-and-error times for finding a right model for an application [10 11]. Moreover, an excessive number of hidden layers or hidden neuron could negatively affect the generalization of MNN that leads to over-fitting issues [12].

Recently, extreme learning machine (ELM) for single-hidden layer feedforward neural networks (SLFNs) has been introduced to the local minimum without complex parameters tuning [13]. The ELM has two parameters to be optimized i.e. the activation function and the number of hidden nodes. The inherent characteristics of simple structure, excellent predictive performance, and high learning speed attribute the superiority of ELM
Both weights and biases for hidden nodes are randomly generated without iteratively adjusting in ELM [14, 15]. Even though this will increase the learning speed and reduce the number of parameters that need to be optimized, the initialization of the hidden layer biases and input weights that done randomly may make ELM unstable in practices [15, 16].

The simplest structure of ANN i.e. single layer ANN that coupled with Levenberg-Marquardt [17] or Gradient Descent technique [18] has shown a potential to achieve faster convergence with lower computational complexity implemented in system identification applications e.g. real-time adaptive control applications and online system identification. Besides, a single linear layer that trained with Widrow-Hoff delta rule has been successfully implemented to model high dimensional NIR spectral data to predict the boiling point of diesel fuel without any data reduction approach and achieved better performance compared with principal component regression (PCR) and Partial least square (PLS) [19]. However, the single layer that trained with Widrow-Hoff delta rule depends on two important factors i.e. the adaption cycle and learning rate. Consequently, a trial and error approach is needed to optimize these two parameters to avoid overfitting problems. Perhaps different algorithm for updating weight that has been proved to improve the performance of the multilayer could further reveal the potential of single linear layer ANN in NIR spectroscopic analysis.

2.5. Genetic Algorithm

The genetic algorithm is developed by Darwin’s principle of the survival of the fitness in order to solve optimization problems (Holland, 1975; Goldberg, 1989). It is used in the medicine and engineering, space research and banking to find the best solution of a different type of problems. The optimal solution is found by a population of competing solutions evolve over time. It is not guaranteed to find the optimum but helps to avoid local maxima. The first is the Genetic algorithm is to have an initial population of randomly generated chromosomes. Each chromosome consists of several genes. The reproduction, selection, crossover, and mutation are genetic operators which help to make a new generation.
Selection of the best chromosomes can be found by rating the individual chromosomes by their associated fitness. Figure 2.3 explains the method of genetic algorithm in a simple way.

![Figure 2.3: Producer of Genetic Algorithm](image)

### 2.5.1 Important Consideration

There is some consideration should be taken to ensure high accuracy of the performance of the genetic algorithm to choose the most important feature of near infrared data.

#### 2.5.1.1 Wavelength Interval Selection

The wavelength selection is an important parameter to design genetic algorithm. For example, if the number of wavelengths increases, the computation will be more complex. To reduce complexity several modifications to have been reported such as pruning of the least relevant regions by some deterministic method, binding of neighboring variables,
spectral windows of varying size. Selection of wavelength interval instead of single wavelengths seems appropriate because NIR absorption bands are broad and smooth. Modification of interval partial least squares (IPLS) was introduced by Leardi and Nørgaard [45]. The method aims to reduce the search space by removing the least informative spectral regions. The dataset that used in the study was polymer film additives and acids in brassica seeds. The result was a reduction in time and comparable accuracies when selection with a previous GA-PLS method [46].

The ranked regions genetic algorithm (RRGA) was introduced by Goicoechea and Olivier [47]. They used RMSEPs IN spectral windows and the intervals are inversely proportional to their fitness, averaged over 100 runs. The result indicates low in computational complexity and inherent ability to avoid over-fitting without explicit monitoring.

2.5.1.2 Termination Stop Criteria

This generational process is repeated until a termination condition has been reached. Common terminating conditions are:
1. A solution is found that satisfies minimum criteria
2. Fixed number of generations reached;
3. Allocated budget (computation time/money) reached;
4. The highest ranking solution’s fitness is reaching or has reached a plateau such that successive iterations no longer produce better results;
5. Manual inspection

It is important to proof convergence because it assures that the solution is optimal. Today the biggest challenge in the implementation of Gas is to decide when to stop the algorithm. Attempts to provide stopping criteria of a GA based on the time the algorithm is being executed and obtained. Time based stopping criteria are mainly of two kinds. The easy way is to decide the upfront the number of iterations to be executed. Second is execution time of the algorithm. This method is requiring a good knowledge about the globally optimal solution, which is not always available.
A new method is to use the underlying fitness function values to calculate the auxiliary values as a measure of the state of the convergence of the GA. Best worst, Phi, Kappa was defined as a convergence measure in.

2.5.2 Genetic Algorithm and Near Infrared Spectroscopy

A genetic algorithm has been widely used to select an important set of explanatory variables for regression analysis. They are being increasingly used for variable selection in diverse areas of chemistry: mass spectrometry, near-infrared (NIR) and mid-infrared spectroscopies, and principal components for multivariate models, etc. Most application of Gas in near infrared spectroscopy contains wavelength selection, for which different approaches have tried. Wavelength selection can improve model accuracy, even in conjunction with latent variable extraction, the robustness of the model against disturbances, reduce the complexity and make the inference of the model easier. In comparative studies, GAs usually outperforms other methods in many aspects. The drawback is the usually longer computation time in comparison to deterministic methods. GAs seems to suffer less from over-learning, probably due to its ability to get stuck in local optima of a noisy landscape, despite the fact that is regarded as a global search method. The cause of over-learning is that measurement noise, which is specific to a data set used to evaluate the fitness function, is used as a predictor variable. Hence, the regression model is over-fitted and its ability to generalize becomes inferior.

2.5.3 Previous Research on Genetic Algorithm with other Regression

There is some Previous Research combine Genetic Algorithm with other types of Regression such as Partial Last Square, Neural Network, and Principle component. From table 2.2 shows that various ways have been used for implementing genetic algorithm with other regressions. The gene is an important parameter for a genetic algorithm which could be variable or interval and for chromosome which consists of numbers of a gene which the length of chromosome depends on the problem that wants to solve. For a crossover,
the operator has different types which could be a single crossover or double crossover and the most used is a single crossover. The mutation percentage is small and normally 1% as shown in this Table 2.2. There are many types of regression that coupled with a genetic algorithm such as PLS, PC, and ANN which proof that genetic algorithm is a good algorithm for feature selection. Table 2.3 shows the number of generations which most of the researchers used in the range of 100 to 200 generation. The root mean square error was mostly used as fitness function as shown in Table 2.3. The selection has many types of methods, but the most used is roulette wheel method.

Table 2.1: Previous Research on Genetic Algorithm with other Regression

<table>
<thead>
<tr>
<th>Chromosomes number</th>
<th>Scheme of encoding</th>
<th>Crossover type</th>
<th>Mutation type</th>
<th>Regression mode</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Percent of total wavelength</td>
<td>either individual wavelengths or wavelength intervals</td>
<td>Single cross over 50%</td>
<td>0.05 percent Mutation</td>
<td>PLS</td>
<td>[48]</td>
</tr>
<tr>
<td>Populations of 30 individuals</td>
<td>Encoded as real values</td>
<td>Single point crossover</td>
<td>N/A</td>
<td>PLS</td>
<td>[49]</td>
</tr>
<tr>
<td>100</td>
<td>Real value encode</td>
<td>Probability for a variable to undergo crossover: 80%</td>
<td>1%</td>
<td>PLS</td>
<td>[50]</td>
</tr>
<tr>
<td>A population of 100 individuals strings,</td>
<td>Binary string</td>
<td>N/A</td>
<td>The mutation rate of 1%</td>
<td>PC</td>
<td>[51]</td>
</tr>
<tr>
<td>30 chromosomes</td>
<td>N/A</td>
<td>N/A</td>
<td>1%</td>
<td>PLS</td>
<td>[52]</td>
</tr>
<tr>
<td>64</td>
<td>Binary strings</td>
<td>Double point</td>
<td>0.005</td>
<td>PLS</td>
<td>[53]</td>
</tr>
<tr>
<td>N/A</td>
<td>The real value</td>
<td>Single point crossover</td>
<td>The rate of 1%</td>
<td>INS</td>
<td>[54]</td>
</tr>
<tr>
<td>200</td>
<td>Binary string</td>
<td>Probability of 0.7 for crossover</td>
<td>Probability Of 0.1</td>
<td>N/A</td>
<td>[55]</td>
</tr>
</tbody>
</table>
Table 2.2: Selection method, Generation, and fitness function of previous research

<table>
<thead>
<tr>
<th>Selection method</th>
<th>Generation</th>
<th>Fitness function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roulette wheel selection method</td>
<td>50 generation stop</td>
<td>Root-mean-square error</td>
<td>[48]</td>
</tr>
<tr>
<td>Stochastic universal sampling (SUS)</td>
<td>100</td>
<td>Root-mean-square error</td>
<td>[49]</td>
</tr>
<tr>
<td>Binary tournament Selection method</td>
<td>200</td>
<td>Root-mean-square error</td>
<td>[50]</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>Press</td>
<td>[51]</td>
</tr>
<tr>
<td>N/A</td>
<td>100</td>
<td>Root-mean-square error</td>
<td>[52]</td>
</tr>
<tr>
<td>N/A</td>
<td>100</td>
<td>Root-mean-square error</td>
<td>[53]</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>ANNs with a fixed architecture are used as fitness functions. The fitness of each individual in this generation is determined by the RMSE of the ANNs</td>
<td>[54]</td>
</tr>
<tr>
<td>Roulette wheel selection method</td>
<td>200</td>
<td>Root-mean-square error</td>
<td>[55]</td>
</tr>
</tbody>
</table>
2.5.4 Neural Network Coupled with Genetic Algorithm

A genetic algorithm has improved the performance of neural network regression model. The study solved the problem of Back propagation algorithm trained NN (BPNN) works with steepest descent principle which has a major limitation in the solutions getting trapped at local minima. Evolutionary GA replaces the back propagation algorithm to searches for global minima at many distinct locations simultaneously in huge space. Thus heuristic search technique (that is, GA), restrict getting trapped at local minima solutions. The GA-NN predictions are found better than those obtained for BPNN and RNN, towards all the process parameters [56]. The improved predictions by GA-NN might be due to the evolutionary search carried out at many distinct locations simultaneously in huge search space. The gradient search method adopted by BPNN and RNN has the greater probability to get stuck at local solutions. Conversely, the evolutionary learning mechanism of GA, initiate a search at many spatial distinct locations. Another recent study coupled the neural network with a generic algorithm using near infrared spectroscopy. The study uses GA-NN in near infrared spectroscopy to predict the ripeness grades of oil palm fresh fruit. GA has successfully reduced the learning time and optimized the network architecture of NN in terms of minimizing the identification error systematically and avoiding the traditional iterative method [57]. Another study use variable selection process based on a modified genetic algorithm with the fixed number of selected variables was proceeded, which can reduce the training time and enhance the predictive ability when coupled with the artificial neural network to determine the Trimethoprim [9]. Another study compares the prediction efficiency of the different type of linear calibration models using near infrared absorbance spectral data of vegetable oils. The applied model types were the PCA-MLR and pls and PCA-ANN and GA-NN. Among all the prediction models, the GA-NN perform the best root mean square error [58].
2.6 Summary

From the literature review, the importance of predicting the diesel fuel properties plays an important role in for the vehicle’s engine. Near infrared spectroscopy is fast inexpensive technique compensate the traditional techniques which are widely used for predicting the diesel fuel properties. Multivariable calibration models improve the performance of near infrared spectroscopy by trying to remove unwanted data such as noise and establish a strong relationship between near infrared spectroscopy and diesel fuel properties. The artificial neural network is an efficient technique that commonly used with near infrared spectroscopy. Multilayer neural network and extreme learning machine are methods to improve the performance of the artificial neural network. The disadvantages of these two methods are the complexity of the model, and the difficulty to determining its parameters such as activation function and a number of hidden neurons. Single layer ANN shows promising in establishing better relationship between a component of interest and NIR spectrum, a different algorithm for updating weight that has been proved to improve the performance of the multilayer could further reveal the potential of single linear layer ANN in NIR spectroscopic analysis. A genetic algorithm is an efficient algorithm widely used as feature selections in near infrared spectroscopy. Artificial neural network coupled with genetic algorithm improves the performance of predicting models.
CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the simulation, analyzing and developing a new predictive model based on single layer neural network to predict the diesel fuel properties using the Matlab software are explained and discussed. First, Section 3.2 theory and algorithm of a single layer, multilayer and learning algorithm such as Levenberg-Marquardt (LM) and Scaled Conjugate Gradient (SCG) are explained and discussed. Second, in section 3.3, an experimental part which includes the efficient of dataset and parameter setting of single layer and multilayer is implemented. Section 3.2 and Section 3.3 used the full dataset without reduction and implemented the proposed model for prediction the diesel fuel properties. Third, in section 3.4, genetic algorithm is used as feature selection and coupled with single layer trained with LM and SCG in order to improve the predictive models and selected the most important wavelengths. Finally, in section 3.4, evaluation of the proposed methods is determined by the root to mean square error (RMSE) and correlation coefficient.

3.2 Flow of the works

Figure 3.1 shows the overall work of the methodology which the first section. Full dataset are used and implemented for single layer and multilayer, and were trained with LM and SCG.
The parameter setting for LM and SCG are settled using MATLAB software and diesel fuel properties are predicted in this section. While Second section single layer trained with LM and SCG and coupled with genetic algorithm for predicting the boiling point of the diesel fuel.

Figure 3.1: Overall works of the methodology
3.3 Theory and Algorithm

In this section, single layer network and multilayer network are presented. The two algorithms which used in both single layer and multilayer neural network are discussed. The proposed method which is based on implementing multilayer algorithm into single layer neural network is explained by given a general and specific description of the implementation of proposed method.

3.3.1 Single Layer Network

The single layer network that has no intermediate layer consists of an input layer and an output layer. Each element of the input vector is connected to each neuron input through the weight matrix. The ith neuron has a summer that gathers its weighted inputs and bias to form its own scalar output. The various scalar output is taken together form an element net input vector. Finally, the neuron layer outputs form a column vector. The derivative of mean square error for a single layer is shown below (Eqn 3.1-3.4) which prove that the first derivative is always minus the input for single layer network

\[
\frac{\partial E^2(K)}{\partial w_{1,j}} = 2 \times E(K) \times \frac{\partial E(k)}{\partial w_{1,j}} \quad (3.1)
\]

\[
\frac{\partial E^2(k)}{(\partial w_{1,j})} = \frac{\partial}{(\partial w_{1,j})} \left[ t(k) - (\sum_{i=1}^{R} w_i^T \times p(K) + b) \right] \quad (3.2)
\]

\[
\frac{\partial E(k)}{\partial w_{1,j}} = -p_j(k) \quad (3.3)
\]

\[
\frac{\partial E^2(K)}{} = -2 \times E(K) \times p(K) \quad (3.4)
\]

Where Kth number of iteration, E is the error, w is the weight and p refer to the input.

A matrix of the first-order partial derivatives of a vector-valued function is called the Jacobian matrix. In the case of a neural network, it is an N-by-W matrix, where N is the number of samples in the training set, and W is the total number of parameters (weights + biases) of the network. Jacobin matrix for a single layer of the mean square error function is stated in Eqn. 3.5.
\[ J = \begin{pmatrix}
\frac{\partial E_1}{\partial w_1} & \cdots & \frac{\partial E_1}{\partial w_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial E_p}{\partial w_1} & \cdots & \frac{\partial E_p}{\partial w_n}
\end{pmatrix} \]  \tag{3.5}

Where \( p \) is the number of input.

The Jacobian matrix is very important for different training algorithms such as Levenberg-Marquardt algorithms because when the performance function has the form of a sum of squares then the Hessian matrix can be approximated as stated in Eqn.3.6.

\[ H = J^T \times J \]  \tag{3.6}

The Jacobin matrix in a single layer is always negative the input which is a help to make the model of neural network fix when applied a single layer with Levenberg-Marquardt.

### 3.3.2 Multilayer Network

Multilayer network models consist of a highly interconnected network of nonlinear processing elements, called hidden neurons. They derive their structure from that of the brain’s biological neurons. Among all the ANNs, the back-propagation algorithm is probably the most widely used supervised learning algorithm in multilayer network models. The networks always include an input layer, at least one hidden layer, and an output layer. First, signals are transmitted to hidden layer from the input layer, and the neurons in hidden layer’s values are renewed by the hidden layer’s transmit function. Then, the signals transmitted to the output layer, where the transmit function of output layer update the neurons values. The first performance of the feedforward phase is on an input pattern to calculate the net error, then, the algorithm uses this computed error of output to change the values of weight in the backward direction. The error is slowly reduced through the hidden layers.
REFERENCES


[39] Li QC, Li DX, Chen QY. “Prediction of pre-oxidation efficiency of refractory


[42] Arora I, Saha A. “Comparison of back propagation training algorithms for software defect prediction”. InContemporary Computing and Informatics (IC3I), 2016 2nd International Conference on 2016 Dec 14 (pp. 51-58). IEEE.


[44] Garg VK, Bansal RK. “Comparison of neural network back propagation algorithms for early detection of sleep disorders”. InComputer Engineering and Applications (ICACEA), 2015 International Conference on Advances in 2015 Mar 19 (pp. 71-75). IEEE.


