

Evaluation of Predictive Capability of Artificial Neural Network and Multiple Linear Regression: A Case Study of Lipid Extraction from Microalgae using Quaternary Solvents

S. I. Mustapha^{1a}, F. A. Aderibigbe^{1b}, and I. A. Mohammed^{1c}

Department of Chemical Engineering, University of Ilorin, Ilorin, Nigeria

^{1a}mustapha.si@unilorin.edu.ng, ^{1b}aderibigbe.fa@unilorin.edu.ng,

^{1c}mohammed.ia@unilorin.edu.ng

Corresponding author: mustapha.si@unilorin.edu.ng ;

Tel: +234 806 718 6215

Abstract

This present study was carried out to evaluate the performance of artificial neural network (ANN) and multiple linear regression (MLR) as modelling tools for predicting the lipid yield obtained from microalgae using quaternary solvent mixtures. The MLR models were developed using two different mathematical softwares viz. Microsoft excel (model 1) and Polymath (model 2). The augmented simplex lattice design under the mixture methodology of the Design Expert software was used to generate the design of experiments used in this study. A comparison of the models developed using ANN and MLR for the extraction process was carried out based on pertinent statistical parameters. Although the results from both MLR models were very close, the calculated values of coefficient of determination (R^2) of 0.9934 and the average absolute deviation (AAD) of 2.4789 for the ANN model when compared with values obtained from the MLR ($R^2 = 0.9898$ and AAD = 2.9821 for model 1) and ($R^2 = 0.9898$ and AAD = 2.9825 for model 2) showed that the ANN model was more accurate and precise than the MLR models. The actual maximum lipid yield of 19.4 wt% lipid g⁻¹ DCW was obtained at solvent mixture in the volume ratio of 1:5:1:1 for chloroform, methanol, ethanol, and dichloromethane respectively.

Keywords:

Artificial neural network, multiple linear regression, lipid extraction, modelling

1. Introduction

The Bligh and Dyer's method using a mixture of chloroform and methanol is known to be the conventional method used for the extraction of lipids from algal biomass (Jeon *et al*, 2013). However, the highly toxic nature of these solvents remain a concern especially for industrial scale application (Iverson *et al*, 2001). To address the challenges of finding new solvent mixtures with less toxicity, the extraction process to investigate various solvent mixture composition such as quaternary solvents can be conducted under optimization studies which can be accomplished with the aid of statistical and mathematical tools (Rizwan *et al*, 2015). Augmented simplex lattice design (ASLD) under the mixture design methodology is one of the statistical tool used to design an experiment involving mixture of components and develop empirical model relating the input variables to the response (Chatsungnoen and Chisti, 2016; Zuorro *et al*, 2016). The optimization procedure based on mixture design includes statistical experimental designs and regression analysis under a set of constrained equations. In general, since theoretical relationships between response variables and causal factors are not clear, multiple regression analysis can be applied to the prediction of response variables on the basis of a second order polynomial equation. Unfortunately, prediction of responses based on the polynomial equation is often limited to low levels, resulting in poor estimation of optimal formulations (Arulsudar *et al*, 2005). In order to overcome the shortcomings in multiple linear regression (MLR), a multi-objective simultaneous optimization technique incorporating an artificial neural network has been developed (Soji-Adekunle *et al*, 2018). Artificial Neural Network (ANN) is a learning system based on a computational technique, which attempts to simulate the function and structure of the biological nervous system (Adetiba and Olugbara, 2015). ANN could be applied to quantifying a non-linear relationship between the studied input variables and the responses by means of iterative training of data obtained from a designed experiment. Hence, the main objective of this study is to compare the prediction accuracy of Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) for lipid yield extracted from microalgae using quaternary solvent mixtures.

In this present study, the design of the experiment generated using augmented simplex lattice design was used to develop models by both ANN and MLR. The MLR was carried out using different mathematical softwares

which include Microsoft excel, and Polymath. The performance of the developed models for extracted lipid yield was compared based on pertinent statistical parameters such as correlation coefficient (R^2), mean average error (MAE), mean square error (MSE), standard error of prediction (SEP), and average absolute deviation (AAD).

2. Methodology

2.1 Design of Experiment (DoE)

The augmented simplex lattice design (ASLD) under the mixture methodology of the Design Expert software (version 11) was used to create the design of experiment (DoE) for the composition of quaternary solvent mixtures for lipid extraction from microalgae. The factors of the studied solvent composition (chloroform, methanol, ethanol, dichloromethane) are each varied from volumes of 0% (0 mL) to 100% (20 mL) based on preliminary experiment carried out to determine the effect of total solvent volume to biomass ratio on the lipid yield recovered from microalgae and the maximum lipid yield was obtained at total solvent volume to biomass ratio of 20 mL:1 g. The ASLD generated a total of 20 solvent compositions and the total sum of the solvent volumes in each mixture was always 100% (20 mL) as shown in Table 1.

The lipid contents from the microalgae were extracted using the conventional solvent extraction method reported by Guldhe *et al* (2014) using known volume of the solvent mixture composition based on the design of experiments as shown in Table 1.

2.2 Multiple Linear Regression (MLR) and Artificial Neural Network (ANN)

The data of the response (lipid yield) obtained from the extraction experiments were fitted to a multiple linear regression model using Microsoft excel (Microsoft office, 2016) and Polymath (Version 6.0) mathematical software. The Polymath (model 2) was fitted to a linear function of the form of equation 1

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_1X_2 + a_6X_1X_3 + a_7X_1X_4 + a_8X_2X_3 + a_9X_2X_4 + a_{10}X_3X_4 \quad (1)$$

where $a_0, a_1, a_2, a_3, a_4, \dots, a_n$ are regression parameters, to a set of N tabulated values of $X_1, X_2, \dots, X_n$ (independent variables) versus Y (dependent variable). It was ensured that the number of data points was greater than $n+1$ (thus $N \geq n+1$). In this case, the number of observation N is 20 while the number of independent variables is 10. The program calculates the coefficients $a_0, a_1, a_2, \dots, a_n$ by minimizing the sum of squares of the deviations between the calculated and the data for Y .

The ANN model was developed using the MATLAB R2015a neural network toolbox (Math Works, Inc., Natick, MA, USA). The data set was divided into three subsets in order to obtain the desired model. For this study, 70% of the data set was used for training, 15% was applied for validation, while the reliability of the model was tested with the remaining 15% of the data set. A feed forward network also known as multi-layered perceptron (MLP) neural network was employed using Levenberg-Marquardt back-propagation algorithm to develop the model. In this work, the hyperbolic tangent sigmoid (tansig) and pure-linear (pure-lin) was used as the transfer function for the input layer to hidden layer mapping and hidden layer to output layer mapping respectively. The number of neurons in the hidden layer of the MLP was determined experimentally by identifying the number of neuron with maximum correlation coefficient value from a range of neuron numbers tested (Adetiba and Olugbara, 2015).

2.3 Statistical Data Analysis

The performance of ANN and MLR models were compared using statistical parameters such as mean square error (MSE), standard error of prediction (SEP), mean average error (MAE), average absolute deviation (AAD), correlation coefficient (R), coefficient of determination (R^2) and adjusted R^2 .

3. Results and Discussion

The predicted values obtained for lipid yield from the ANN and

MLR models are presented in Table 1. The predicted lipid yield from ANN model was in the range of 5.75% to 20.11%. The optimum lipid yield from the use of the quaternary solvent mixture as predicted by ANN model was obtained under the same mixture composition as with the predicted MLR models (12.5% chloroform, 62.5% methanol, 12,5% ethanol and 12.5% dichloromethane) as shown in run 12.

Table 1: Experimental and predicted lipid yield values by ANN and MLR

Run	Composition of components % (mL)				Actual	ANN	Lipi	
	d yield (% lipid g ⁻¹ DCW)						MLR	
	Chloroform (X ₁)	Methanol (X ₂)	Ethanol (X ₃)	DCM (X ₄)			Model 1	Model 2
1	100 (20)	0 (0)	0 (0)	0 (0)	8.20	7.5964	7.4786	7.4786
2	50 (10)	50 (10)	0 (0)	0 (0)	20.20	20.1842	19.7876	19.7873
3	50 (10)	0 (0)	50 (10)	0 (0)	13.80	13.9149	13.7439	13.7441
4	50 (10)	0 (0)	0 (0)	50 (10)	7.80	7.7004	7.2795	7.2796
5	0 (0)	100 (20)	0 (0)	0 (0)	17.60	17.2591	17.3086	17.3089
6	0 (0)	50 (10)	50 (10)	0 (0)	15.40	15.4268	15.6044	15.6046
7	0 (0)	50 (10)	0 (0)	50 (10)	18.00	18.9652	17.8197	17.8197
8	0 (0)	0 (0)	100 (20)	0 (0)	6.40	6.3078	6.4872	6.4873
9	0 (0)	0 (0)	50 (10)	50 (10)	12.00	12.0127	12.1760	12.1765
10	0 (0)	0 (0)	0 (0)	100 (20)	6.00	5.4877	5.6948	5.6948
11	62.5 (12.5)	12.5 (2.5)	12.5 (2.5)	12.5 (2.5)	12.80	12.8207	14.0072	14.0072
12	12.5 (2.5)	62.5 (12.5)	12.5 (2.5)	12.5 (2.5)	19.40	20.0584	19.5649	19.5650
13	12.5 (2.5)	12.5 (2.5)	62.5 (12.5)	12.5 (2.5)	15.20	14.9780	13.9377	13.9379
14	12.5 (2.5)	12.5 (2.5)	12.5 (2.5)	62.5 (12.5)	12.40	12.3756	12.6773	12.6775
15	25 (5)	25 (5)	25 (5)	25 (5)	16.80	17.9453	16.9816	16.9817
16	0 (0)	100 (20)	0 (0)	0 (0)	17.00	17.2591	17.3086	17.3089
17	0 (0)	0 (0)	100 (20)	0 (0)	6.20	6.3078	6.4872	6.4873
18	100 (20)	0 (0)	0 (0)	0 (0)	7.00	7.5964	7.4786	7.4786
19	0 (0)	0 (0)	0 (0)	100 (20)	5.40	5.4877	5.6948	5.6948
20	50 (10)	0 (0)	0 (0)	50 (10)	7.20	7.7004	7.2795	7.2796

3.1 Multiple Linear Regression Modelling

The linear model obtained from MLR modelling of the experimental data via model 1 and model 2 are shown in equation 2 and 3 respectively. The mathematical expressions describes the relation between the independent variables (solvent composition) and the

response (predicted lipid yield) for MLR models as shown on Table 1.

$$\%Yield = +7.48 + 0.492X_2 - 0.0496X_3 - 0.0892X_4 + 0.0739X_1X_2 + 0.0676X_1X_3 + 0.0069X_1X_4 + 0.0371X_2X_3 + 0.0632X_2X_4 + 0.0609X_3X_4 \quad (2)$$

$$\%Yield = +0.3739X_1 + 0.8654X_2 + 0.3244X_3 + 0.2847X_4 + 0.0739X_1X_2 + 0.0676X_1X_3 + 0.0069X_1X_4 + 0.0371X_2X_3 + 0.0632X_2X_4 + 0.0609X_3X_4 \quad (3)$$

Where X_1 , X_2 , X_3 , and X_4 are the composition of chloroform, methanol, ethanol and dichloromethane, respectively. From model 1 (equation 2), the model equation shows that model term X_2 (Methanol) has a positive and highest effect on the lipid yield with coefficient 0.492 while the model terms X_3 (Ethanol) and X_4 (Dichloromethane) have negative effects on the lipid yield. Furthermore, the model shows that all the interactions (X_1X_2 , X_1X_3 , X_1X_4 , X_2X_3 , X_2X_4 , and X_3X_4) have positive effects on the lipid yield. This implies that the use of any of the solvent mixture is well favoured toward the lipid yield statistically with the interaction between chloroform and methanol having the highest effect (coefficients 0.0739) and the interaction between chloroform and dichloromethane with the lowest effect (coefficient 0.0069). Similar trend was observed with the model 2 (equation 3) except that in this case all the model terms have positive effects on the lipid and the term X_1 was also introduced in the model 2. The summary of the analysis of variance for the multiple linear regression for model 1 and model 2 are presented in Table 2 and Table 3 respectively. As shown in Table 3, all the terms are significant at 95% confidence level for the model 2 unlike the case of model 1 (Table 2) where some of the model terms are not significant although they may have effect on the lipid yield.

Table 2: The results of ANOVA for multiple linear regression model 1

		<i>Degree of freedom</i>	<i>Sum of Squares</i>	<i>Mean Square</i>	<i>F-value</i>	<i>Significance F</i>			
Regression		10	481.8246	48.18246	108.2986	3.87E-08			
	Residual	10	48.0494338						
	Total	20	486.768						
<i>Source</i>	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
Intercept	7.478625	0.4901	15.24108	3E-08	6.385303	8.571948	6.385303	8.571948	
X ₁	0	0	65535	-	0	0	0	0	
	X ₂	0.491514	0.034	14.17724	-	0.414266	0.568762	0.414266	0.568762
	X ₃	-0.04957	0.034	-1.42971	0.18329	-0.12681	0.027681	-0.12681	0.027681
	X ₄	-0.08919	0.034	-2.57205	0.027793	-0.16645	-0.01193	-0.16645	-0.01193
	X ₁ X ₂	0.073935	0.007	10.0415	1.53E-06	0.05753	0.090341	0.05753	0.090341
	X ₁ X ₃	0.067611	0.007	9.18256	3.45E-06	0.051205	0.084017	0.051205	0.084017
	X ₁ X ₄	0.006928	0.01	1.172395	0.268211	-0.00624	0.020096	-0.00624	0.020096
	X ₂ X ₃	0.037065	0.007	5.025779	0.000517	0.020633	0.053498	0.020633	0.053498
	X ₂ X ₄	0.063179	0.007	8.580571	6.34E-06	0.046773	0.079584	0.046773	0.079584
	X ₃ X ₄	0.060854	0.007	8.264893	8.84E-06	0.044448	0.07726	0.044448	0.07726

Table 3: The results of ANOVA for multiple linear regression model 2

Variable	Coefficients	95% confidence		Comments
	0.3739	0.05		Significant
X ₂	0.8654	0.05		Significant
X ₃	0.3244	0.05		Significant
X ₄	0.2847	0.05		Significant
X1 X ₂	0.0739	0.016		Significant
X1 X ₃	0.0676	0.016		Significant
X1 X ₄	0.0069	0.013		Significant
X ₂ X ₃	0.0371	0.016		Significant
X ₂ X ₄	0.0632	0.016		Significant
X3 X4	0.0609			
	0.016	Significant		Variance = 0.4943
Number of independent variables = 10				RMSD = 0.1112
Regression not including a free parameter				Number of observations = 20

3.2 Artificial Neural Network Modelling

The optimal number of neurons in the hidden layer of the multi-layered perceptron (MLP) neural network was determined experimentally as illustrated in Figure 1. From the plot, it can be seen that the highest correlation coefficient (R) was obtained at hidden layer number of neuron 10.

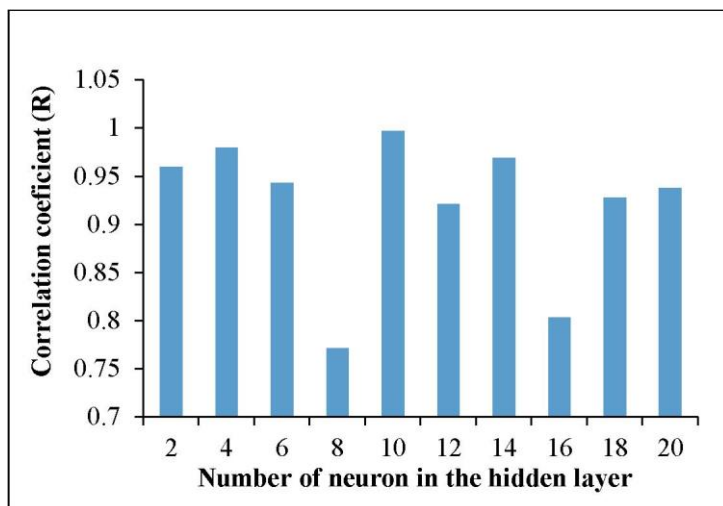


Figure 1. Optimal number of hidden layer neurons

Hence, the ANN architecture assembled contain an input layer with four neurons (the composition of quaternary solvent), a hidden layer with 10 neuron and an output layer with one neuron (lipid yield) as shown in Figure 2. The regression plots showing the coefficient of correlation for training, testing, validation and whole data sets are depicted in Figure 3.

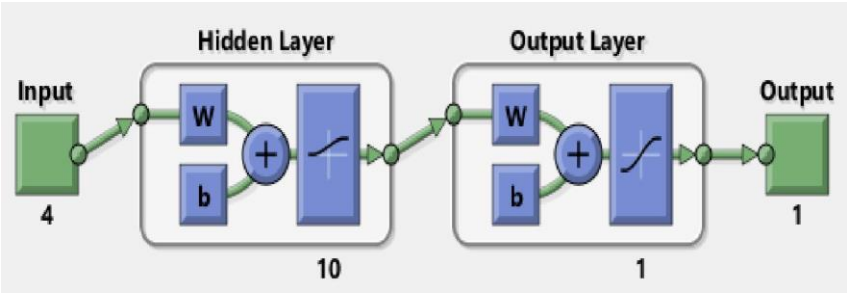


Figure 2. Architecture of multi-layered perceptron neural network

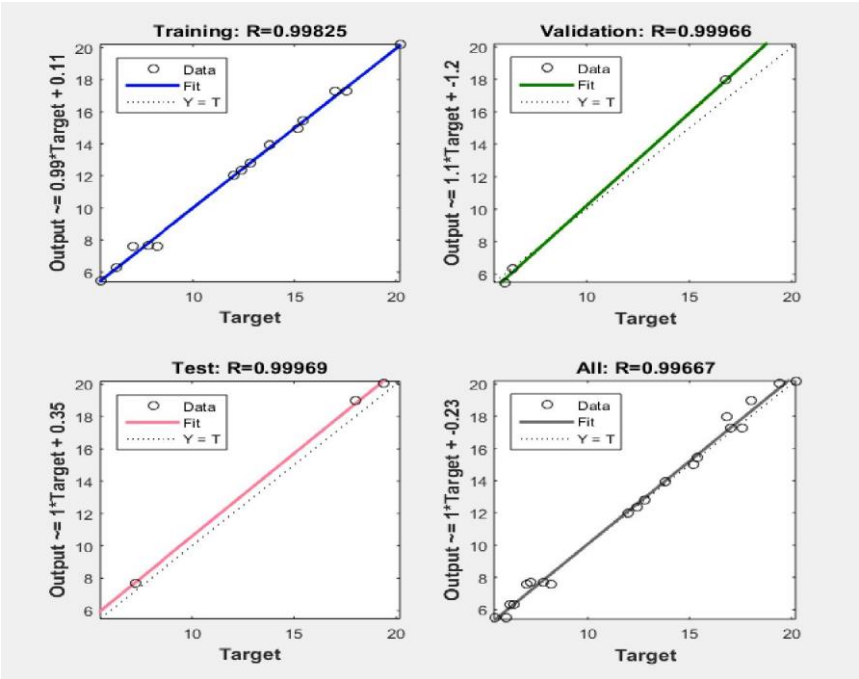


Figure 3 Regression plots for the ANN model

216189

2689

3.3 Evaluation of the Predictive Capability of ANN and MLR

The evaluation of the predictive capabilities of ANN and MLR models for lipid yield was carried out and the results were compared as shown in Table 4. Statistical parameters such as MSE, MAE, SEP, AAD, R and R² were used in comparing the performance of the developed models. The results obtained shows that the ANN model performed better than the MLR models, although the performance of both MLR models were very close and satisfactory. The MSE, MAE, and SEP values indicated that lower errors were obtained with the ANN model compared to the MLR models. The average absolute deviation of the ANN model was also found to be lower compared to the MLR models. More so, the ANN model had higher precision and accuracy as shown with the values of R (0.9967), R² (0.9934) and adjusted R² (0.9875) compared with R (0.9949), R² (0.9898) and adjusted R² (0.8807) for model 1 and R (0.9949), R² (0.9898) and adjusted R² (0.9807) for model 2 obtained from the MLR.

Table 4: Predictive evaluation of ANN and MLR models

Parameter	ANN	MLR	
		Model 1	Model 2
RMSE	0.4181	0.4538	0.4538
MAE	0.2669	0.3124	0.3124
SEP	3.4158	3.7079	3.7079
AAD	2.4789	2.9821	2.9825
R	0.9967	0.9949	0.9949
R ²	0.9934	0.9898	0.9898
Adjusted R ²	0.9875	0.8807	0.9807

The plot of experimental and predicted lipid yield values against the experimental run is shown in Figure 4. The figure also confirms that the predicted values for the ANN model were much more closely aligned with the experimental values than the predicted values of the MLR models.

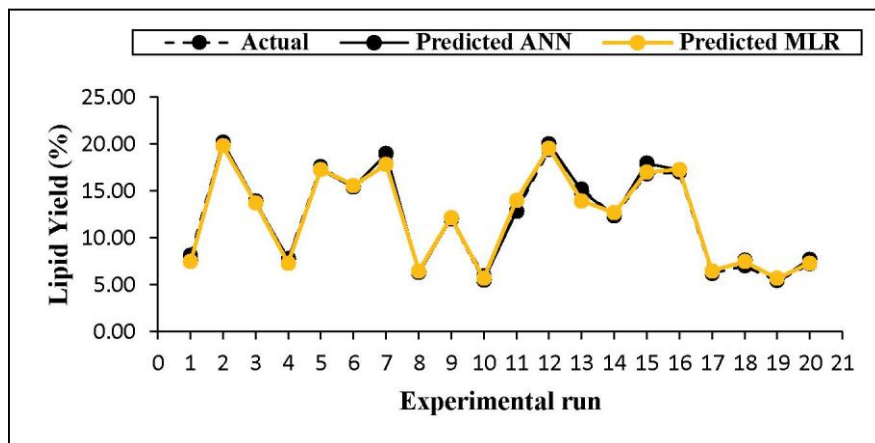


Figure 4 Experimental and predicted lipid yield from ANN and MLR against experimental run

4. Conclusion

This work assessed the performance of artificial neural network (ANN) and multiple linear regression (MLR) models in predicting the lipid yield from microalgae using quaternary solvent mixtures. Augmented simplex lattice design under the mixture methodology was employed to generate the experimental data used for developing the models. The statistical indicators used for comparing the predictive capabilities of the developed models showed the performance of the ANN model ($R^2 = 0.9934$ and AAD = 2.4789) to be superior to the MLR models ($R^2 = 0.9898$ and AAD = 2.9821 for model 1) and ($R^2 = 0.9898$ and AAD = 2.9825 for model 2). The results showed all the developed models performed adequately and the performance of the MLR models (model 1 and 2) was found to be very close.

References

Adetiba, E. and Olugbara, O. O. (2015). Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. *The Scientific World Journal*, 2015: 1 - 17.

216391

2692

Arulsudar, N.; Subramanian, N. and Murthy, R. (2005). Comparison of artificial neural network and multiple linear regression in the optimization of formulation parameters of leuprolide acetate loaded liposomes. *J Pharm Pharm Sci*, 8(2): 243.

Chatsungnoen, T. and Chisti, Y. (2016). Optimization of oil extraction from *Nannochloropsis salina* biomass paste. *Algal Research*, 15: 100-109.

Guldhe, A.; Singh, B.; Rawat, I.; Ramluckan, K. and Bux, F. (2014). Efficacy of drying and cell disruption techniques on lipid recovery from microalgae for biodiesel production. *Fuel*, 128: 46-52.

Iverson, S. J.; Lang, S. L. and Cooper, M. H. (2001). Comparison of the Bligh and Dyer and Folch methods for total lipid determination in a broad range of marine tissue. *Lipids*, 36(11): 1283-1287.

Jeon, J.-m.; Choi, H.-W.; Yoo, G.-C.; Choi, Y.-K.; Choi, K.-Y.; Park, H.-Y.; Park, S.-H.; Kim, Y.-G.; Kim, H. J. and Lee, S. H. (2013). New mixture composition of organic solvents for efficient extraction of lipids from *Chlorella vulgaris*. *Biomass and Bioenergy*, 59: 279-284.

Rizwan, M.; Lee, J. H. and Gani, R. (2015). Optimal design of microalgae-based biorefinery: Economics, opportunities and challenges. *Applied Energy*, 150: 69-79.

Soji-Adekunle, A. R.; Asere, A. A.; Ishola, N. B.; Oloko-Oba, I. M. and Betiku, E. (2018). Modeling of synthesis of waste cooking oil methyl esters by artificial neural network and response surface methodology. *International Journal of Ambient Energy*(just-accepted): 1-34.

Zuorro, A.; Miglietta, S.; Familiari, G. and Lavecchia, R. (2016). Enhanced lipid recovery from *Nannochloropsis* microalgae by treatment with optimized cell wall degrading enzyme mixtures. *Bioresource technology*, 212: 35-41.