Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 4 PAGE: 1195-1211 YEAR: 2023
DOI:10.17798/bitlisfen.1362136

# Theoretical Models Constructed by Artificial Intelligence Algorithms for Enhanced Lipid Production: Decision Support Tools

Aytun ONAY[1*]

[1]*Turkish Aeronautical Assocation University Engineering Faculty, Software Engineering, Ankara, 06790, Turkey*

*(ORCID: 0000-0001-5104-0668)*

**Keywords:** Artificial intelligence algorithms, Biodiesel, Chlorella minutissima, Ensemble learning algorithms, Microalgal lipid content, Response surface methodology.

**Abstract**

Theoretical models that predict the lipid content of microalgae are an important tool for increasing lipid productivity. In this study, response surface methodology (RSM), RSM combined with artificial neural network (ANN), and RSM combined with ensemble learning algorithms (ELA) for regression were used to calculate the maximum lipid percentage (%) from Chlorella minutissima (C. minutissima). We defined one set of rules to achieve the highest lipid content and used trees.RandomTree (tRT) to simulate the process parameters under various conditions. Among the various models, results showed the optimum values of the root mean squared error (0.2156), mean absolute error (0.1167), and correlation coefficient (0.9961) in the tRT model. RSM combined with tRT estimated that the lipid percentage was 30.3% in wastewater (< 35%), lysozyme (≥ 3.5 U/mL), and chitinase (< 15 U/mL) concentrations, achieving the best model based on experimental data. The optimal values of wastewater concentration, chitinase, and lysozyme were 20% (v/v), 5 U/mL, and 10 U/mL, respectively. Also, the if-then rules obtained from tRT were also used to test the process parameters. The tRT model served as a powerful tool to obtain maximum lipid content. The final rankings of the performance of various algorithms were determined. Furthermore, the models developed can be used by the fuel industry to achieve cost-effective, large-scale production of lipid content and biodiesel.

## 1. Introduction

The demand for renewable energy has increased in emerging nations. The majority of countries are trying to find renewable energy alternatives to replace fossil fuels [1], [2]. Biodiesel production has become very important for the global fuel market due to the consumption of fossil fuels, growing energy consumption, climate change, and environmental pollution, but it needs highly innovative research and development [3], [4]. Biodiesel obtained from microalgal biomass is an encouraging source of renewable energy; however, there are some difficulties in producing algae-derived biodiesel [5]-[7]. Therefore, researchers are looking into new economical production techniques [8]-[10]. Biomass features intracellular lipids and rigid cell walls thus an appropriately selected solvent is needed to extract the lipids [11], [12]. Many studies have been conducted to investigate cell wall destruction and optimal solvent selection to improve microalgal lipid yields. Water in biomass also acts as a barrier to achieving lipid extraction. The biodiesel efficiency of microalgae is based on dry biomass [13], [14]. The industry has struggled to decrease the cost of not only wet but also high-water-content algal biomass. Microalgae are used as health foods and in cosmetics by different industries. They have produced some impressive products, like lipids, carbohydrates, pigments, and proteins [15]. The high-volume

*Corresponding author: aonay@thk.edu.tr
*Received: 18.09.2023, Accepted: 02.12.2023*

production of various metabolic products looks quite promising for obtaining lipids, biomass, green energy products, enzymes, polymers, toxins, and pigments. Microalgae use solar energy and consume carbon dioxide for photosynthesis [16], [17]. The cultivation conditions are important for microalgae, affecting their growth characteristics and cellular composition [18], [19]. The most-used culture media include photoautotrophic, heterotrophic, and mixotrophic media. The major drawback is the cultivation cost to obtain biomass. Nitrogen (80 kg) and phosphorous (5 kg) are needed to get about one ton of biomass [20]. Therefore, optimizing the process parameters is important to obtain high amounts of lipids in a cost-effective way. Researchers have searched for different theoretical methods and developed software tools to predict the maximum lipid content and biomass productivity based on cultivation conditions and medium compositions. Among them, RSM is the major optimization technique [21]–[24]. RSM combined with ANN has also been performed to determine the optimal medium components and cultivation conditions [25], [26]. Furthermore, based on optimization studies in the literature, the genetic algorithm with support vector machines (SVM) and fuzzy logic methods (FLM) were implemented. Mondal et al. [27] performed the central composite design (CCD) to optimize three main process parameters: temperature, light intensity, and CO2. Chlorella sp. BTA 9031 was used in the cultivation medium for biomass and lipid formation. The optimum levels of 28.26 °C, 76.64 μmol m−2s−1, and 4.00% CO2 were determined for temperature, light intensity, and CO2, respectively. The results show that the model achieved reliable results. Alam et al. [28] searched for the effects of process parameters on three-phase partitioning performance (TPP). The temperature, incubation duration, and extraction time were optimized by using RSM. C. vulgaris, which has a high water content, was used for the extraction of lipids. At a ratio of 1:1 solvent to DKP, the optimum results for the temperature, incubation duration, and extraction time were 60 °C, 120 min, and 60 min, respectively. The results revealed that the lipid extraction procedure can be performed to form biodiesel. Ishola et al. [29] focused on developing a model. The transformation of sorrel (Hibiscus sabdariffa) oil to H. sabdariffa methyl esters (HSME) was modeled using RSM, ANN, and a neuro-fuzzy inference system (ANFIS). According to the statistical tests, the models developed to describe the transesterification process were very precise and reliable. The ANFIS model achieved the highest R2 value (0.9944). The ANFIS model optimized with genetic algorithm (GA) predicted the optimum conditions for maximum HSME product yield (99.71 wt%) at the methanol/oil molar ratio of 8:1. Catalyst weight, reaction time, and temperature were found to be 1.23 wt%, 43 min, and 65 °C, respectively. The results show that the HSME can be used as an alternative to petro-diesel.

In this study, RSM, RSM combined with ANN, and RSM combined with ELA [random forest (RF), random tree (RT), and bagging (BA)] were used to calculate the maximum lipid percentage (%) from C. minutissima using the MATLAB, SIMULINK and Weka software [30]-[34]. Different concentrations of wastewater, chitinase, and lysozyme were tested using Box-Wilson design to observe the process parameters that cause changes in the lipid content.

Wastewater can be used to produce biofuels from microalgae. High levels of carbon, nitrogen, and phosphorus in wastewater can act as nutrients for microalgae and enable them to grow. The microalgal mass obtained in large amounts can also be used in biofuel production. Thus, economically advantageous biofuel can be obtained without providing carbon, nitrogen, and phosphorus input to the system from the outside [35]. Chitinase and lysozyme are two important enzymes that can take part in the degradation of the outer surface of microalgae cells. With this pre-treatment technique, the outer membranes of microalgae can be further broken down, and quickly, enabling greater access to their metabolic contents. In this case, it means producing more lipids [36]. Since the optimization of variables such as carbon source, temperature, pH, and light in biodiesel production is well known, the parameters used in this study can increase the lipid amount and generate more metabolic content. The synergistic effects of the parameters on lipid content were also examined by performing a circumscribed central composite design (CCCD) of RSM. In this study, AI algorithms were run together with RSM in a combined structure. The literature describes the ANN and RSM models on biofuel production as “black boxes.” By developing explainable platforms and simplifying the models for the researcher’s understanding, we make the models more robust and transparent. Also, there are few reported models and articles on biofuel research using combined RSM-ELA models. The trees.RandomTree (tRT) model is an explainable machine-learning model that lays out one set of rules to achieve the optimal parameters and serves as a decision-support tool to achieve maximum lipid content. This theoretically obtained decision mechanism is of great importance to researchers. Interactive response surface modeling

(rstool) and prediction plots were also developed and drawn for the three process parameters of wastewater concentration (v/v %), chitinase (U/mL), and lysozyme (U/mL). The novelty of this study is the production of lipids, which are biodiesel raw materials, theoretically in high amounts by using wastewater-containing nutrients for microalgae and enzyme technologies such as chitinase and lysozyme. This theoretical study can be achieved by using the parameters without the need for repeated experiments in the production of biodiesel. Thus, the cost will be reduced, as no experimental repetitions are made. The paper also overviews AI algorithms that can be used in other projects with similar conditions and practical applications.

## 2. Material and Method

ANN is ideal for building regression models and conditions, and it has the ability to learn and model nonlinear and complicated interactions, which is crucial because the parameter relationships (wastewater concentration, chitinase, and lysozyme) between inputs and outputs are nonlinear and complex. As the number of parameters used in biofuel studies increases, using a decision tree structure in optimization studies is very important in terms of simulating parameter values under various environmental conditions. In this regard, our research is important. Bagging of regression trees helps to make machine-learning algorithms work better and more accurately. They are used to deal with trade-offs between bias and variance and to lower the variance of a prediction model. Therefore, RSM combined with ELA models were used for regression. The performance results of all the algorithms were given to the researchers for comparison.

### 2.1. Cultivation of Microalgae and Lipid Extraction

The experimental data related to C. minutissima CCALA 723 were obtained from Van YYU, Department of Environmental Engineering, Turkey. Microalgae were grown in BG-11 medium and municipal wastewater. Wastewater was prepared according to Khan's method with a few modifications [37]. The medium includes glucose (540 mg/L), ammonium chloride (180 mg/L), sodium bicarbonate (140 mg/L), and potassium dihydrogen phosphate (55 mg/L). The medium was mixed with distilled water in a ratio of 1:50. Wastewater (0–100%) was mixed with BG-11 medium proportionally. Chitinase (0–50 U/mL) and lysozyme (0–20 U/mL) were combined for the disruption of microalgae. Microalgae were harvested by centrifugation at 3000g for 10 min, and lipid extraction was performed according to the Folch method [38]. The final ratio of methanol, chloroform, and water was 1:1:0.9. Lipid samples were weighed gravimetrically.

### 2.2. Design of Experiments

#### 2.2.1. Quadratic Response Surfaces for Modeling

The synergistic and individual effects of the parameters were examined with CCCD to calibrate the quadratic models and construct a five-level three-factor model. This model generated 24 theoretical runs. The theoretical design contained 8 factorials, 6 axial points, and 10 replicate results at the center points, maintaining knowledge of the interior of the experimental region. To determine the repeatability of the method, the center point was repeated ten times. Using a Box-Wilson model, the ideal levels of three process variables were determined: wastewater concentration (% v/v) ($X_1$), chitinase (U/mL) ($X_2$), and lysozyme (U/mL) ($X_3$). Factors $X_1$, $X_2$, and $X_3$ were figured out on -2 (the axial points), -1, 0 (central point), +1 and +2 (the axial points) theoretical levels, as given in Table 1. Each test was implemented on the data to determine the average of the duplicates. The relationship between the independent factors and the dependent value Y was modeled as the second-degree polynomial in $X_1$, $X_2$, and $X_3$. The second-order polynomial coefficients were solved by MATLAB to obtain high amounts of lipids. The results were also used to generate the response surface graphs. The quadratic equation was solved to determine the response Y value of the model given as follows:

$$\mathbf{Y}_{\text{lipidpercentage}}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1}X_1+\boldsymbol{\beta_2}X_2+\boldsymbol{\beta_3}X_3+\boldsymbol{\beta_{12}}X_1X_2+\boldsymbol{\beta_{13}}X_1X_3+\boldsymbol{\beta_{23}}X_2X_3+\boldsymbol{\beta_{11}}X_1X_1+\boldsymbol{\beta_{22}}X_2X_2+\boldsymbol{\beta_{33}}X_3X_3 \quad (1)$$

where Y was the predicted lipid percentage (%). $\boldsymbol{\beta_0}$ represents the intercept; $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}$ the linear effect, $\boldsymbol{\beta_{11}}, \boldsymbol{\beta_{22}}, \boldsymbol{\beta_{33}}$ the quadratic coefficients, and

**Table 1.** The axial and central points for the model

| Factors | Variables codes | Coded factor levels* | | | | |
|---|---|---|---|---|---|---|
| | | -2 (-∝) | -1 | 0 | +1 | +2 (+∝) |
| Wastewater concentration (% v/v) | $X_1$ | 0 | 20 | 50 | 80 | 100 |
| Chitinase (U/mL ) | $X_2$ | 0 | 5 | 10 | 25 | 50 |
| Lysozyme (U/mL) | $X_3$ | 0 | 2 | 5 | 10 | 20 |

*+∝ and - ∝ indicate the axial points

$\boldsymbol{\beta_{12}\beta_{13}\beta_{23}}$ the interaction effect. The theoretical data obtained from CCCD were adjusted for the quadratic equation.

### 2.2.2. ANOVA Statistics

The analysis of variance (ANOVA) was calculated to determine the accuracy of the RSM model with the degree of freedom (DF) values for each term, sum of squares (SS) for the regression model, mean squared error (MSE) for each term, F-statistic value, and P-value for the F-test. The coefficient of determination ($R^2$) and the adjusted $R^2$ were used to denote the validity of the model. The $R^2$ value is advanced by approximately 1. When p has a very small value ($p < 0.05$), the involved term in the model achieves an important action on the response. The best model is constructed by getting a very good agreement between predicted and experimental results. MATLAB was run to perform the ANOVA calculation.

### 2.3. Application of the ANN Model

ANN closely mimics the human brain and has several advantages over the traditional theoretical model. It can be implemented quickly into rigorous phenomenological models. The network design was constructed with MATLAB to model the optimization process of lipid extraction. A two-layer feed-forward network with sigmoid hidden neurons and linear output neurons was defined for the network architecture. The Levenberg-Marquardt algorithm was used to train the neural network. The experimental data were arbitrarily split into three groups: 70% for the training set, 15% for the validation set, and 15% for the test set. The model was based on this dataset. After training the network, MSE, and regression analysis were used to evaluate its performance. MSE measures the average squared difference between the output and targets. The lower the MSE value, the stronger the model is. In addition, regression (R) predicts the relationships between the response variable and predictors or features. An R is expected to be close to 1. In this study, the network was formed from an input layer, a hidden layer, and an output layer. An input layer had three neurons, such as wastewater concentration, chitinase, and lysozyme. The number of hidden neurons (20) was optimized to get the minimum value of MSE. Also, an output layer had one neuron.

### 2.4. Interactive Response Surface Modeling and Prediction Plots

The maximum values of y were searched by using rstool. Rstool displays a graphical user interface for exploring one-dimensional contours of multidimensional response surface models interactively using MATLAB. Syntax was determined as rstool (x, y, quadratic, alpha, xname, yname). X defines the input values, y defines the response values, and alpha is 0.01. The y values were theoretically found for the ANN and ELA models, and the y values for the RSM model were experimentally determined. For the fitted response surface, rstool plots two red curves, representing a 95% confidence band. Input values were given in the text boxes on the horizontal axis and marked by vertical dashed blue lines in the plots. In this study, interactions and full quadratic models were used. The interaction model contains constant, linear, and interaction terms. The full quadratic model also contains squared terms. These terms take place in the quadratic response surface equation as given in Eq. 1. By interacting with the rstool interface, the RSM, ANN, and ELA demos are able to display the outcomes of simulated experiments involving data or pre-specify combinations of predictors. The maximum lipid percentages for each model were predicted. Root mean squared error (RMSE) and beta/constant (intercept) values were used to evaluate the model performance.

### 2.5. Ensemble Learning Algorithms

RF, RT, and BA were used for building regression models. Weka software was used to execute algorithms to obtain the maximum lipid content. The experiments were done in 24 sets, as mentioned above.

To improve the accuracy of the model, runs 15 to 24 were set at the same input conditions. After

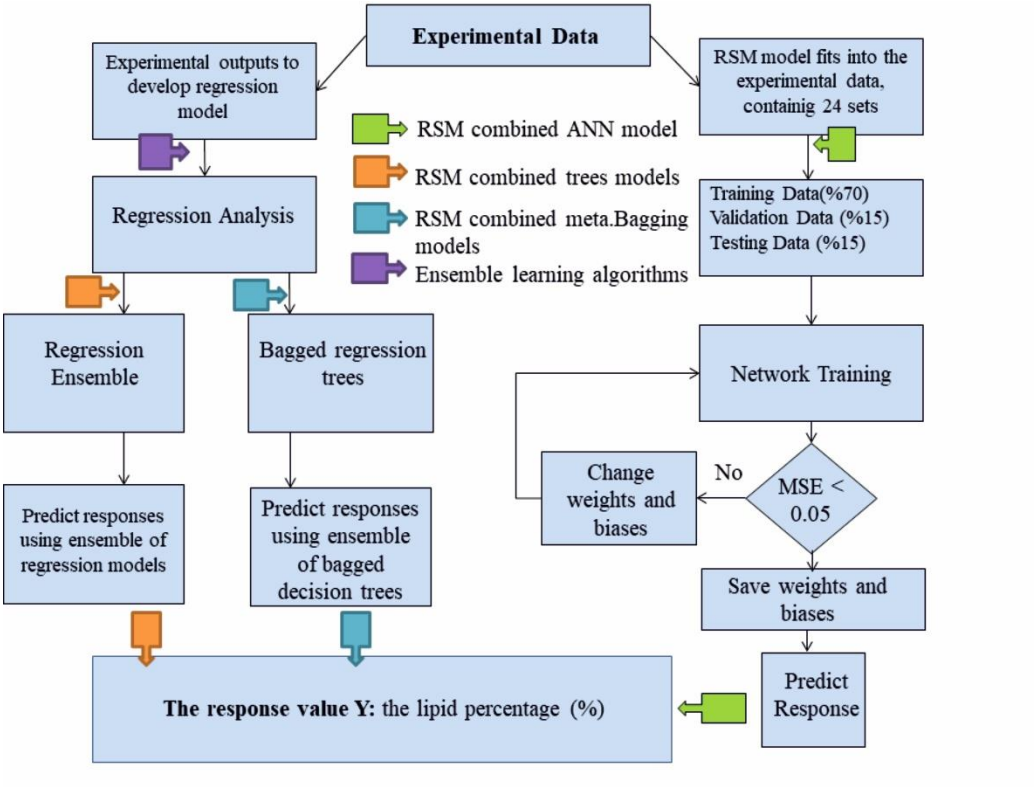


**Figure 1.** Graphical representation of the various models.

performing the experiments, the outputs were collected. Then, the RSM model was obtained to fit the data. The experimental outputs were used to develop regression models. Predictor variables ($X_1$, $X_2$, $X_3$) and response variables (Y) built the model. The correlation coefficient, mean absolute error, and RMSE were calculated to prove the accuracy of the regression model parameters. Figure 1 shows the different models in a simplified diagram.

AI algorithms are used to perform various tasks such as classification, clustering, and regression. In this study, RF and RT were used to determine the most effective parameters for lipid production. The model developed with RF or RT can be used as a decision-support tool. RF was extended by Leo Breiman and Adele Cutler, who obtained a trademark for the tool [39], [40]. Both RF and RT achieved good results to estimate the lipid percentage (%). The if-then rules obtained from tree algorithms are provided in the results section.

The learning models used by Random Forest consist of numerous decision trees. For each model, we create a sample dataset by randomly selecting rows and features from the dataset. This section is known as Bootstrap.

BA has three stages. Firstly, it creates several samples from the main training set by assigning each combination the same probability. Then, it performs a base learner (regression tree in our case) model for the samples. Finally, it averages the values calculated for the test sets over the models. Weka builds this process using "meta-learning" under the name bagging. The dataset was in .csv format and contained 24 experimental run results with regard to the three attributes (inputs) and response values (outputs; see Table 2). Our data matrix was composed of three parameters. The training set in .csv format was used to obtain developed regression models in this study. After the file in .csv format is converted to .arff, this model can be loaded into the Weka software and used to make predictions on new data. All models obtained in the study can be run by researchers to perform the prediction procedure on the lipid percentage (%).

Because of the little available data (Table 2), cross-validation was performed to improve the model's predictive performance. K-fold (a cross-validation technique) was used in this part of the study, and 10-fold validation (k = 10) was chosen. This technique is done k times such that each subset is validated exactly once. For smaller datasets, k-fold technique is recommended. Because training and validation are performed multiple times, cross-

validation can be a computationally costly procedure. It is, nonetheless, an important stage in the model building since it reduces the risk of overfitting or underfitting a model. The average errors across all k = 10 partitions are reported in this study. After loading the models to Weka software, the results of the new test sets for which the option "Re-evaluate model on current test set" is worked can be obtained using the algorithms above. The predictions for each new test are given in the "Classifier Output" pane in the program.

## 3. Result and Discussion

In this section, the interactions of wastewater concentration, chitinase, and lysozyme and their impacts, positive or negative, on the lipid percentage (%) are discussed.

### 3.1. Comparison of The Quadratic RSM and ANN Model

RSM is a practical technique for searching for the effects of several factors affecting the process of calling the optimal terms. This method provides a way to reduce the number of experiments, evolve statistical interpretation possibilities, and specify the interaction among various variables. The equation was improved to correlate Y with the functions of $X_1$, $X_2$, and $X_3$. The empirical model developed by RSM is given by the equation below:

$$Y = 20.879 + 0.0897\ X_1 - 0.0927\ X_2 + 1.446\ X_3 + 0.0018\ X_1X_2 + 0.0016\ X_1X_3 - 0.0031\ X_2X_3 - 0.0014\ X_1X_1 - 0.0011\ X_2X_2 - 0.0604\ X_3X_3 \quad (2)$$

RSM results showed that the MSE, $R^2$, and adjusted $R^2$ values are 0.2414, 0.9765, and 0.9614, respectively. The MSE of a model relative to a test set is the mean of the squared prediction errors in all instances in the test set. When there are no errors in a model, the MSE equals 0. The proportion of the variance in the dependent variable that can be defined by the independent variable is evaluated by $R^2$. $R^2$ indicates how well the data fit the regression model (the goodness of fit). The greater the $R^2$ value, the smaller the MSE. If the value of $R^2$ reaches 1 (ideal-world scenario), the model fully fits the data with an MSE of 0. Adjusted $R^2$ is a modified form of $R^2$ that adjusts for nonsignificant factors in a regression model. The modified $R^2$ indicates whether the addition of more predictors improves a regression model. Adjusted $R^2$ will never exceed $R^2$. $R^2$ states that every variable explains the variation in the dependent variable. Adjusted $R^2$ signifies how much variation is explained by independent variables that affect the dependent variable. A higher adjusted $R^2$ in a model with additional input variables implies that the additional input variables provide value to the model. Table 2 shows both the experimental design and the results.

Lipid content ranges from 21.4% to 30.3% for experimental data (Table 2). Interdependent effects are denoted by a positive sign in front of the variables, whereas opposite effects are denoted as a negative sign in equation Y, which included linear, interaction, and squared terms with the three factors. P-value (probability value) is expected to be less than 0.05, and F-value is expected to be high (~1) in the models. MATLAB was used to construct the models. The maximum lipid percentage of 30.4% was obtained by using RSM, in which wastewater concentration, chitinase, and lysozyme were 20% (v/v), 5 U/mL, and 10 U/mL, respectively. The highest lipid percentage determined by the experiments was 30.3%. ANOVA was used to calculate degrees of freedom (DF), sum of square (SS), MSE, F-statistic value, and P-value for the model. It measures the model effectiveness and the fitness of the regression model. ANOVA results for the RSM are presented in Table 3. As shown in Table 3, the P-value and F-value were found to be 0.00001 and 64.7132, respectively. The results show that the quadratic model is highly important to predict lipid content based on the experimental data.

The regression equations (Eq. 2) that describe the impacts of independent process factors on lipid production are graphically represented in the 3D and 2D plots. The elliptical shape in the contours defines important impacts among parameters. The spherical form of the contour plot indicates that the interaction effect of parameters is insignificant. The maximum value for the lipid content is indicated by the center ellipse in the contour plot. This value is found within the intervals of parameters that have been specified. The 3D and 2D plots that are obtained depend on two parameters, while the third parameter stays the same at the origin. In Figure 2a, a surface plot (3D) displays both mutual and interaction effects between $X_1$ and $X_2$. The increase in $X_2$ from low to high levels decreased the lipid content. The maximum lipid content occurred when $X_1$ (50 %) was at the middle level and $X_2$ was at the minimum level. Lipid content was thus significantly affected by the interaction between $X_1$ and $X_2$. P-value (0.0048333) is an indicator of this. The form of the contours in Figure 2b clearly reveals significant interactions between factors. As shown in Figure 2c, the 3D surface plot displays the mutual and interaction effects of $X_1$ and $X_3$ on the lipid content. Increasing $X_1$ and $X_3$ caused an increase in the lipid content to some extent, but then it had an adverse effect on the

**Table 2.** RSM and ANN model-based design for optimization of lipid content

| Exp. order | Process parameters | | | Lipid percentage (%) | | |
|---|---|---|---|---|---|---|
| | Wastewater concentration (% v/v) | Chitinase (U/mL ) | Lysozyme (U/mL) | Experimental | RSM predicted | ANN predicted |
| 1 | 20 | 5 | 2 | 24.4 | 24.5 | 24.4 |
| 2 | 20 | 5 | 10 | 30.3 | 30.4 | 30.3 |
| 3 | 20 | 25 | 2 | 22.0 | 22.6 | 22.0 |
| 4 | 20 | 25 | 10 | 28.1 | 28.0 | 28.1 |
| 5 | 80 | 5 | 2 | 21.4 | 22.1 | 22.8 |
| 6 | 80 | 5 | 10 | 29.0 | 28.7 | 29.0 |
| 7 | 80 | 25 | 2 | 22.2 | 22.4 | 22.2 |
| 8 | 80 | 25 | 10 | 28.2 | 28.6 | 28.2 |
| 9 | 0 | 10 | 5 | 25.6 | 25.4 | 24.5 |
| 10 | 100 | 10 | 5 | 23.2 | 22.8 | 23.2 |
| 11 | 50 | 0 | 5 | 27.4 | 27.9 | 27.4 |
| 12 | 50 | 50 | 5 | 24.6 | 24.4 | 20.2 |
| 13 | 50 | 10 | 0 | 22.5 | 21.7 | 22.5 |
| 14 | 50 | 10 | 20 | 27.4 | 27.4 | 27.4 |
| 15 | 50 | 10 | 5 | 28.1 | 27.7 | 27.7 |
| 16 | 50 | 10 | 5 | 27.3 | 27.7 | 27.7 |
| 17 | 50 | 10 | 5 | 27.9 | 27.7 | 27.7 |
| 18 | 50 | 10 | 5 | 27.3 | 27.7 | 27.7 |
| 19 | 50 | 10 | 5 | 27.8 | 27.7 | 27.7 |
| 20 | 50 | 10 | 5 | 28.4 | 27.7 | 27.7 |
| 21 | 50 | 10 | 5 | 27.5 | 27.7 | 27.7 |
| 22 | 50 | 10 | 5 | 27.8 | 27.7 | 27.7 |
| 23 | 50 | 10 | 5 | 27.6 | 27.7 | 27.7 |
| 24 | 50 | 10 | 5 | 27.5 | 27.7 | 27.7 |

**Table 3.** ANOVA Results for the RSM

| Source | Lipid percentage (%) | | | | | |
|---|---|---|---|---|---|---|
| | DF [a] | SS[b] | MS[c] | *F*-value | *P*-value | Significance |
| Model | 9.0000 | 140.5901 | 15.6211 | 64.7132 | 0.0000 | Significant |
| Residual | 14.0000 | 3.3795 | 0.2414 | | | |
| Total | 23.0000 | 143.9696 | | | | |

[a]Degrees of freedom for each term
[b]Sum of square for the regression model
[c]Mean squared error for each term

the lipid content. Figure 2d shows that there is no significant interaction effect between $X_1$ and $X_3$ to contribute to lipid content. This can be also observed from the P-value (0.28883).

Figure 2a and Figure 3a show that the interaction of $X_1$ and $X_2$ was more significant than that of $X_2$ and $X_3$ on lipid content. Lipid content also shows a nonlinear effect with the increase of $X_3$ from 0 to 20, and with the increase of $X_2$ from 0 to 50 under constant X1. The irregular contour plot in Figure 3b shows that both $X_2$ and $X_3$ had an interaction effect, but it was not enough to change the lipid content (P-value, 0.4727). According to the ANOVA results, the overall model is important (P-value < 0.00001). ANN, based on feed-forward neural networks with TANSIG transfer function, was used to optimize the process variables on lipid content. The model predicted lipid percentage (%) at the times (t = 0, t + n), presented in Table 2. ANN was iterated with varying numbers of hidden neurons with transfer functions in the layer to optimize the training and validation data set to achieve a coefficient of determination. Figure 4 shows all input and output variables that compare the $R^2$ value of the training, the validation and test data, and all prediction networks, with the values 0.99548, 0.96336, 0.98257, and 0.92838, respectively.

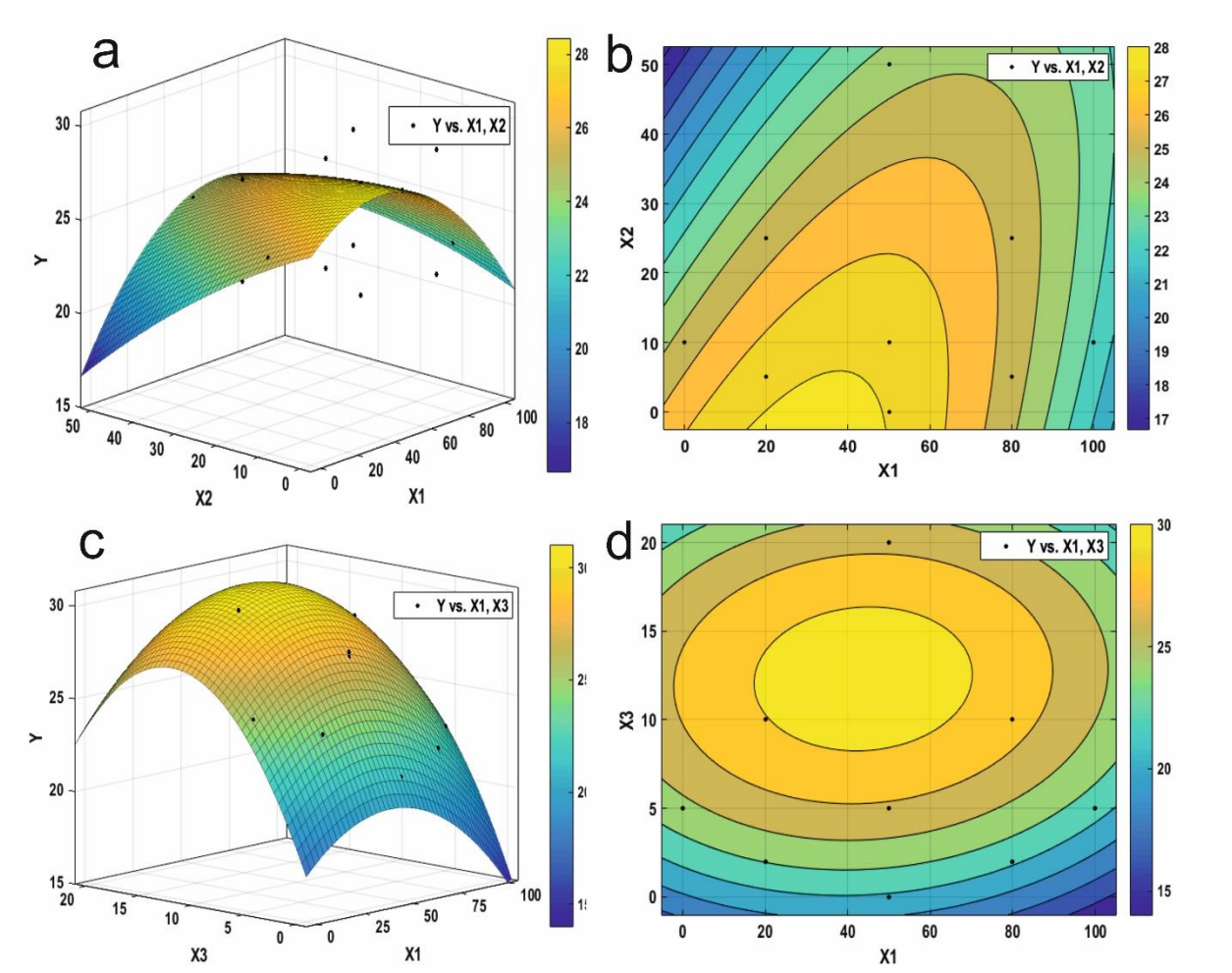


**Figure 2.** 3D and related 2D plots for the various effects of (a-b) wastewater concentration ($X_1$) and chitinase ($X_2$); (c-d) wastewater concentration ($X_1$) and lysozyme ($X_3$) on the lipid percentage (%).

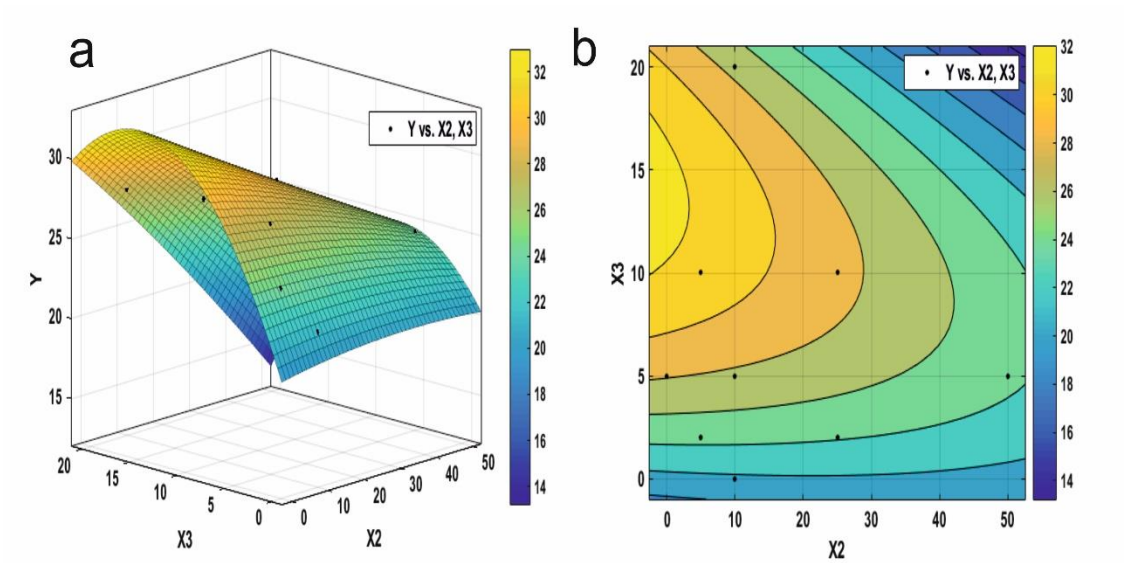


**Figure 3.** (a) 3D and (b) related 2D contour plots for the various effects of chitinase ($X_2$) and lysozyme ($X_3$) on the lipid percentage (%).

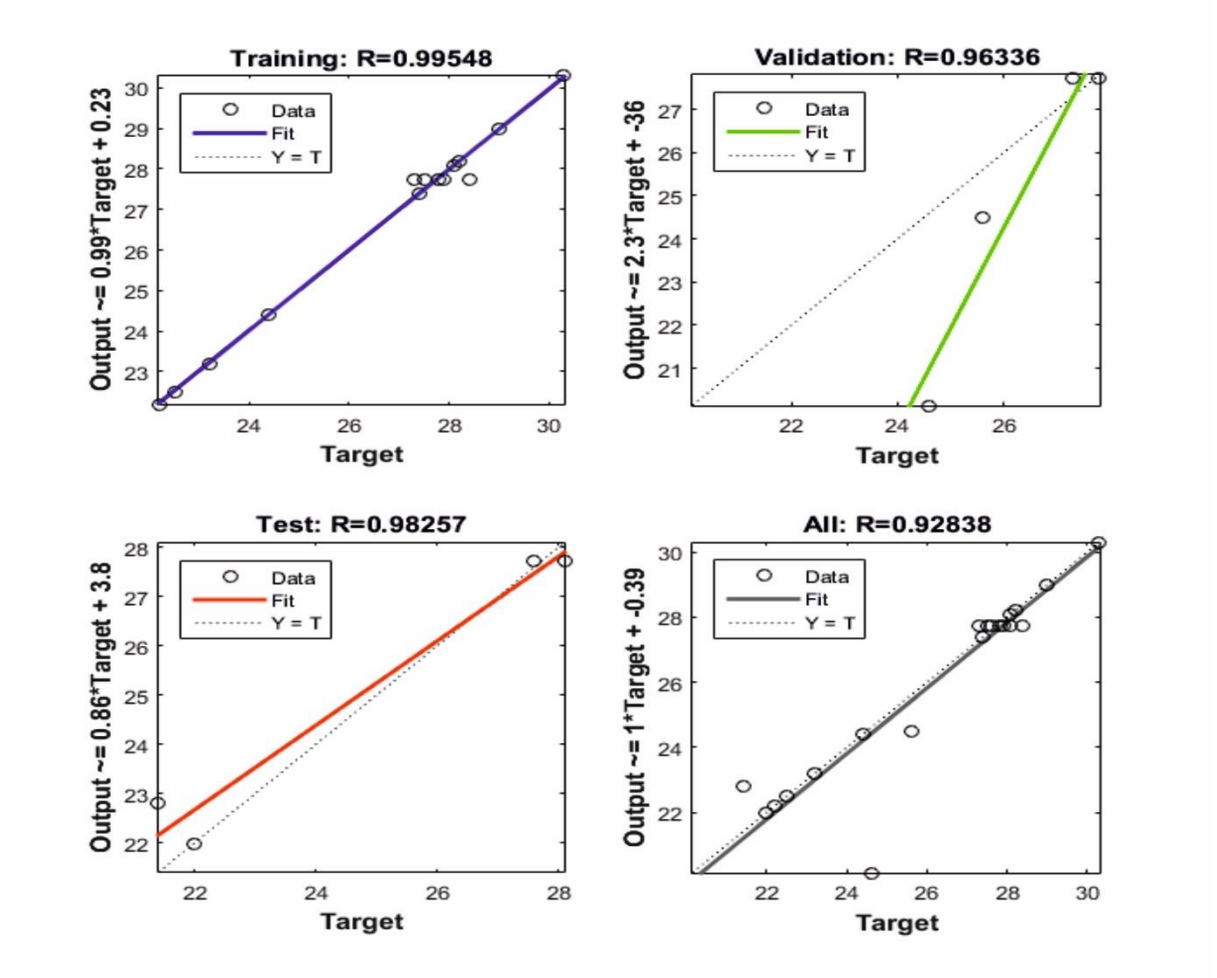


**Figure 4.** Fitness of the different datasets and all prediction networks in the ANN model.

ANN model reached an $R^2$ value close to 1 for the data sets in Figure 4 and had high-accuracy results in modeling to the output parameters estimating lipid content. The MSEs of the ANN model for the training and test data were 0.0484 and 0.5221, respectively. The model estimated the optimum conditions for maximum lipid content as $X_1$ = 20% (v/v), $X_2$ = 5 U/mL, and $X_3$ = 10 U/mL, with a maximum lipid percentage of 30.3%. These results indicate that the ANN outperformed RSM in predicting lipid content.

In this section, literature-based studies on experimental designs and theoretical models are presented. Mäkelä [22] presented a review of the energy field, which provides many aspects of experimental design and RSM with applications. Mäkelä (2017) reported many solutions for the modeling studies. Tourang et al. [20] studied the optimization of macronutrient concentrations for *Spirulina* microalgae, and CCD was used to optimize the culture medium. Muthuraj et al. [41] studied lipid-rich biomass with a high density of *Chlorella sp*. FC2 IITG under photoautotrophic conditions using a process engineering strategy. They used ANN or RSM with a genetic algorithm (GA) for medium optimization, and the results showed that ANN-GA achieved an increase in biomass titer of 157% (0.95 g $L^{-1}$) in a shake flask. Thanaa et al. [42] researched 13 filamentous fungi for their lipid production. CCD was performed to obtain the optimum concentrations to increase lipid productivity, and they concluded *P. brevicompactum* NRC 829 may be used for commercial development. Dammak et al. [43] optimized the V2-strain to maximize the cell growth and lipid content of the oleaginous microalgae by using RSM. The results indicated that strain V2 was convenient for food and nutraceutical applications. By using RSM, Onay (2020) studied increasing the lipid content of *N. gaditana* using several assisted lipid extraction methods. The greatest lipid content was achieved in 10 KCI% osmotic shock, 30 kHz ultrasound, and 10 U/mL lysozyme with a lipid percentage of 37% [44]. Khaouane et al. [45] used a hybrid method with CDD, ANN, and the particle swarm optimization algorithm (PSO) to optimize the culture conditions of pleuromutilin production. The hybrid technique determined the optimum levels of culture conditions at 242 rpm agitation speed,

temperature of 26.88 °C, and pH of 6.06 with a pleuromutilin yield of 10.074 ± 500 μg/g. Sohedein et al. (2020) used RSM to construct an optimization model to extract the maximum biomass (6.25 g $L^{-1}$) and lipid content (14.88% DW) from *Aurantiochytrium sp*. They concluded that biomass concentration in the bioreactor increased 2.12-fold compared to the shake flask culture. The optimized model could provide large-scale biodiesel production [46]. Chakravarty et al. (2019) performed three sets of optimization with a central composite rotary design. Using the effects of nitrate, phosphate, and NaCl, they considered the fourth factor, incubation period, to achieve optimum lipid content in a minimum of days. *Selenastrum sp.* GA66 was subjected to stress factors. Researchers observed that the optimized process parameters required less nitrate. The improved model can be used to obtain biodiesel economically [47]. Zhang et al. [48] improved lipid extraction yield from *Scenedesmus sp.* This study involved cell-wall disruption. The RSM method was used to optimize enzyme-assisted lipid extraction methods. The analysis showed that the optimal conditions produced twice as much lipid content as not using enzymes. Ayoola et al. [26] performed RSM and ANN to determine the effects of KOH and NaOH catalysts on waste groundnut oil (WGO) biodiesel production. They reported that the maximum yield of biodiesel was obtained from KOH catalyst. The results show that ANN and RSM can be used to solve complex problems in biodiesel production. To maximize the lipid content of various microalgae, RSM or RSM combined with ANN methods are the most studied in literature. RSM combined with ELA was also performed in our study for lipid-rich biomass.

### 3.2. Interactive Response Surface Method

Interactive response surfaces with full quadratic and interaction models were improved to calculate the maximum lipid content. Fitted coefficients were determined for RSM, ANN, and ELA-based designs. The RMSE values for the full quadratic models were 0.4913, 0.2500, and 0.6104 for RSM, ANN, and tRF, respectively. When modifying the value of a predictor, all plots are automatically updated to reflect the new data point within the predictor space. Therefore, when the value of a predictor changes, we can simultaneously observe it graphically. The models predicted the maximum lipid percentage at 50% v/v of $X_1$, 12.9167 U/mL of $X_2$, and 5.75 U/mL of $X_3$. The fitted coefficients are displayed in Table 4.

Beta terms were also given to all developed models. From the beta terms, the effects of contributions from linear, interaction, and squared terms on the lipid content (Y), as functions of process variables, were found separately. Rstool models are displayed in Figure 5 for RSM.

The RMSE value of 0.4021 was observed with a lipid percentage of 28.1678 ± 0.85206 % for tRT. In terms of making reliable predictions, a smaller RMSE is better. The experimental results were quite consistent with the predicted results. According to the tRT model, the contribution of linear and interaction terms to equation Y is mostly positive, whereas the contribution of squared terms to equation Y is negative. The lowest RMSE of 0.2500 was obtained from the ANN model. The tRT model also achieved a lower RMSE of 0.4021 than the other decision tree ensemble models.

### 3.3. Ensemble Learning Algorithms for Regression

In the Weka software, optimization of factor levels that satisfy the conditions was performed on the process parameters and responses. Table 5 shows the decision tree ensemble models used for the regression to predict the percentage of lipid content.

According to the tRT model, the optimal values of $X_1$, $X_2$, and $X_3$ were determined at 20 % (v/v), 5 U/mL, and 10 U/mL, respectively. The model estimated a maximum lipid percentage of 30.3%, which is the same as the experimental result. In addition, the highest lipid percentages obtained for tRF and meta.Bagging.Classifier:RandomForest were 29.1% and 29.0%, respectively. RMSE is used to determine the data that is closest to the line of best fit. In this study, RMSE was used in regression analysis to validate experimental data. The predictive accuracy for ensemble algorithms is given in Table 6.

**Table 4.** Fitted coefficients of developed RSM, ANN, and tRF models

| Fitted coefficients | RSM | | ANN | | tRF | |
|---|---|---|---|---|---|---|
| | Full quadratic | Interactions | Full quadratic | Interactions | Full quadratic | Interactions |
| RMSE | 0.4913 | 2.1402 | 0.2500 | 2.0694 | 0.6104 | 1.8028 |
| Lipid percentage (%) | 28.1678 +/- 1.0411 | 26.4028 +/- 2.2909 | 28.2956 +/- 0.5298 | 26.2335 +/- 2.2152 | 27.8616 +/- 1.2935 | 26.5171 +/- 1.9298 |
| ***beta*** | | | | | | |
| Constant | 20.8791 | 28.6302 | 19.7759 | 28.5802 | 21.4511 | 27.4311 |
| $X_1$ | 0.0897 | -0.0523 | 0.1201 | -0.0284 | 0.0797 | -0.0292 |
| $X_2$ | -0.0927 | -0.2249 | -0.0138 | -0.2987 | -0.0710 | -0.1450 |
| $X_3$ | 1.4460 | 0.1138 | 1.4012 | 0.1118 | 1.2457 | 0.1488 |
| $X_1X_2$ | 0.0018 | 0.0018 | 0.0013 | 0.0013 | 0.0009 | 0.0009 |
| $X_1X_3$ | 0.0016 | 0.0015 | 7.6747 | 7.6747 | 0.0004 | 0.0004 |
| $X_2X_3$ | -0.0031 | 0.0134 | 0.0012 | 0.0196 | -0.0014 | 0.0116 |
| $X_1X_1$ | -0.0014 | | -0.0015 | | -0.0011 | |
| $X_2X_2$ | -0.0011 | | -0.0041 | | -0.0002 | |
| $X_3X_3$ | -0.0604 | | -0.0573 | | -0.0499 | |

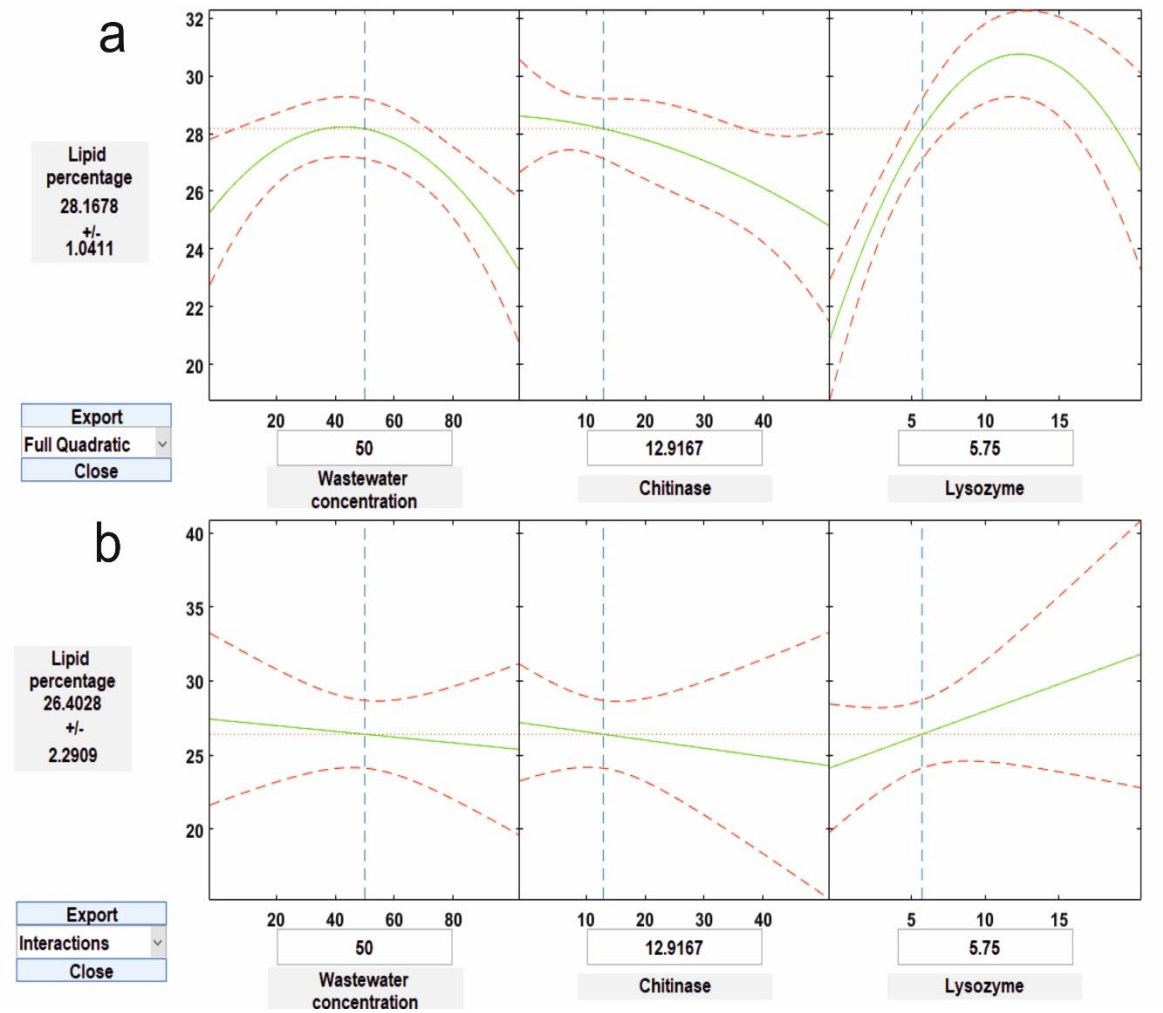


**Figure 5.** Rstool modeling for RSM

**Table 5.** ELA for regression to predict the percentage of lipid content

| Exp. order | Process P[a] | | | Lipid P (%)[b] | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | trees. | | meta.Bagging | |
| | WC (% v/v) [c] | C (U/mL)[d] | L (U/mL)[e] | Exp[f] | RandomForest predicted | RandomTree predicted | Classifier: RandomForest predicted | Classifier: RandomTree predicted |
| 1 | 20 | 5 | 2 | 24.4 | 24.0 | 24.4 | 24.4 | 23.9 |
| 2 | 20 | 5 | 10 | 30.3 | 29.1 | 30.3 | 29.0 | 30.0 |
| 3 | 20 | 25 | 2 | 22.0 | 22.7 | 22.0 | 23.1 | 22.6 |
| 4 | 20 | 25 | 10 | 28.1 | 28.0 | 28.1 | 27.4 | 28.3 |
| 5 | 80 | 5 | 2 | 21.4 | 22.5 | 21.4 | 23.5 | 22.6 |
| 6 | 80 | 5 | 10 | 29.0 | 28.0 | 29.0 | 28.0 | 28.8 |
| 7 | 80 | 25 | 2 | 22.2 | 22.6 | 22.2 | 22.8 | 22.3 |
| 8 | 80 | 25 | 10 | 28.2 | 27.7 | 28.2 | 27.0 | 28.4 |
| 9 | 0 | 10 | 5 | 25.6 | 26.5 | 25.6 | 26.8 | 26.6 |
| 10 | 100 | 10 | 5 | 23.2 | 24.6 | 23.2 | 24.8 | 23.0 |
| 11 | 50 | 0 | 5 | 27.4 | 27.6 | 27.4 | 27.7 | 27.9 |
| 12 | 50 | 50 | 5 | 24.6 | 25.9 | 24.6 | 26.0 | 26.0 |
| 13 | 50 | 10 | 0 | 22.5 | 22.7 | 22.5 | 23.9 | 23.2 |
| 14 | 50 | 10 | 20 | 27.4 | 27.8 | 27.4 | 28.0 | 28.0 |
| 15 | 50 | 10 | 5 | 28.1 | 27.7 | 27.7 | 27.7 | 27.7 |
| 16 | 50 | 10 | 5 | 27.3 | 27.7 | 27.7 | 27.7 | 27.7 |
| 17 | 50 | 10 | 5 | 27.9 | 27.7 | 27.7 | 27.7 | 27.7 |
| 18 | 50 | 10 | 5 | 27.3 | 27.7 | 27.7 | 27.7 | 27.7 |
| 19 | 50 | 10 | 5 | 27.8 | 27.7 | 27.7 | 27.7 | 27.7 |
| 20 | 50 | 10 | 5 | 28.4 | 27.7 | 27.7 | 27.7 | 27.7 |
| 21 | 50 | 10 | 5 | 27.5 | 27.7 | 27.7 | 27.7 | 27.7 |
| 22 | 50 | 10 | 5 | 27.8 | 27.7 | 27.7 | 27.7 | 27.7 |
| 23 | 50 | 10 | 5 | 27.6 | 27.7 | 27.7 | 27.7 | 27.7 |
| 24 | 50 | 10 | 5 | 27.5 | 27.7 | 27.7 | 27.7 | 27.7 |

[a]Process parameters [b]Lipid percentage (%) [c]Wastewater concentration (% v/v) [d]Chitinase (U/mL ) [e]Lysozyme (U/mL) [f]Experimental

**Table 6.** Comparison of the predictive accuracy for ELA models

| Lipid percentage (%) predicted (10-fold validation (k = 10), Total number of instances 24) | | | | | |
|---|---|---|---|---|---|
| Ensemble algorithms | Correlation coefficient | Mean absolute error | Root mean squared error | Relative absolute error (%) | Root relative squared error (%) |
| trees. RandomForest | 0.9731 | 0.5290 | 0.6723 | 25.1221 | 27.4493 |
| trees. RandomTree | 0.9961 | 0.1167 | 0.2156 | 5.5409 | 8.8043 |
| meta.Bagging Classifier: RandomForest | 0.9641 | 0.7202 | 0.9132 | 34.2064 | 37.2844 |
| meta.Bagging Classifier: RandomTree | 0.9795 | 0.4393 | 0.5620 | 20.8628 | 22.9469 |

The statistical analysis showed that the tRT model produced the lowest RMSE value. Additionally, the meta.Bagging.Classifier:RandomTree model has a very low RMSE value. Correlation analysis is performed to evaluate the linear association between two variables, and it takes a value between -1 and +1. In Table 6, tRT achieved the highest correlation coefficient of 0.9961 and the lowest MAE, another loss function performed for regression models. MAE was calculated to prove the sum of the absolute differences between our target and predicted parameters. Therefore, the tRT model was more successful than others. All models can be used to evaluate estimation procedures for the lipid percentage (%). Figure 6 shows the performance of

the RSM, ANN, and tRT models against the experimental data.

Different methods and algorithms have been used in the literature to perform optimization. Zhang et al. [49] developed models for microbial lipid production. The genetic algorithm with SVM was performed to obtain the optimum parameters, including biomass, lipid content, and chemical oxygen demand removal rate. SVM was the best-fitting model to optimize fermentation conditions. The optimum biomass and lipid content were 11.87 g/L and 2.18 g/L, respectively. Nassef et al. [50] focused on enhancing lipid extraction from *S. quadricauda* by using fuzzy modeling and particle swarm optimization. The effect of the process parameters, including power (W), heating time (min), and extraction time (h), on the lipid content was investigated. They reported that the fuzzy model outperformed experimental studies in terms of lipid extraction by 22%.

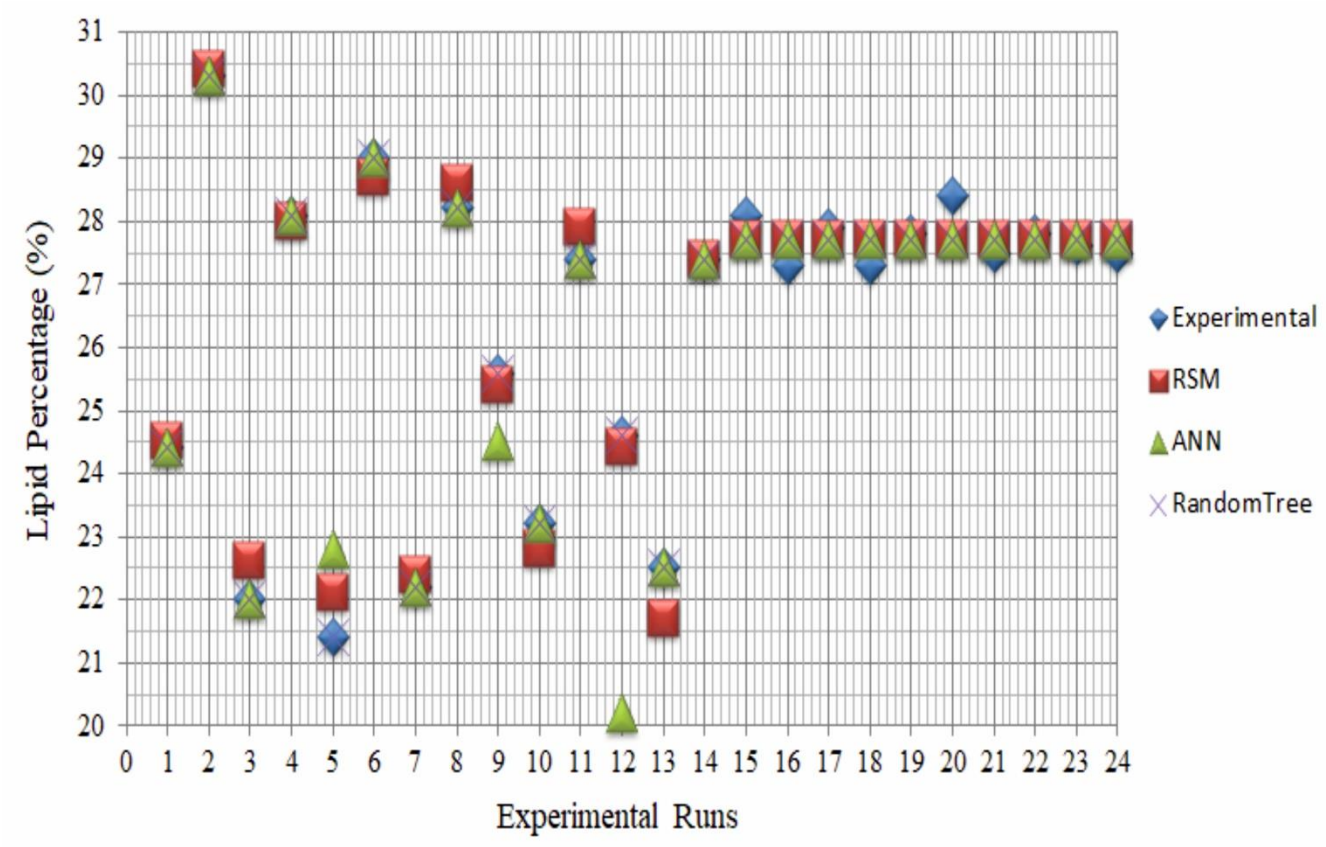


**Figure 6.** Performance of RSM, ANN, and tRT models against experimental data.

### 3.4. Explainable Machine Learning Model: tRT

Large-scale biofuel production from microalgae has not yet been realized. Major research gaps, such as the reduction of energy input, maximization of yield, and those related to the effective use of materials and energy, remain unresolved. In microalgae cultivation, the nutrient supply has a considerable effect on cost, sustainability, and yield [51]. The aim is to research and create simulation tools in order to address these problems. Figure 7 shows the tRT model and the size of the tree.

The ruleset given below was obtained from the tRT model for the dataset (Table 2). It was simulated under various parameters affecting lipid content.

- **if** wastewater concentration < 65
- **and** lysozyme ≥ 3.5
- **and** chitinase < 37.5
- **and** wastewater concentration ≥ 10
- **and** wastewater concentration < 35
- **and** chitinase < 15
- **then** lipid percentage = 30.3

The explainable system [52] can be turned into an application program for different laboratories to use. The user can enter input values into the program and calculate the maximum biodiesel production for each AI model. The biggest advantage of this model is that when human-computer interaction is considered, the structure of the system can be observed in the decision tree. In this way, researchers understand how the decision stage is reached. The decision support tool can be easily used for the following purposes:

- Model validation in labs
- Validation of the lipid-production process in small-scale facilities
- There is no need for repeated experiments on biodiesel production.

In addition, there are many large and small-scale algal biofuel companies producing algae biomass and biofuel throughout the world. Some of the companies throughout the world are Albemarle Corporation (USA), Algenol Biofuels (USA), Aquaflow Bionomics (New Zealand), and Malaysian Rubber (Australia). In the near future, various models and tools can be collected and then presented to these companies as a service (i.e., the AI-Library) to upscale the production volumes.

```
WASTEWATER < 65
|   LYSOZYME < 3.5
|   |   WASTEWATER < 35
|   |   |   CHITINASE < 15 : 24.4 (1/0)
|   |   |   CHITINASE >= 15 : 22 (1/0)
|   |   WASTEWATER >= 35 : 22.5 (1/0)
|   LYSOZYME >= 3.5
|   |   CHITINASE < 37.5
|   |   |   WASTEWATER < 10 : 25.6 (1/0)
|   |   |   WASTEWATER >= 10
|   |   |   |   WASTEWATER < 35
|   |   |   |   |   CHITINASE < 15 : 30.3 (1/0)
|   |   |   |   |   CHITINASE >= 15 : 28.1 (1/0)
|   |   |   |   WASTEWATER >= 35
|   |   |   |   |   LYSOZYME < 12.5
|   |   |   |   |   |   CHITINASE < 5 : 27.4 (1/0)
|   |   |   |   |   |   CHITINASE >= 5 : 27.72 (10/0.11)
|   |   |   |   |   LYSOZYME >= 12.5 : 27.4 (1/0)
|   |   CHITINASE >= 37.5 : 24.6 (1/0)
WASTEWATER >= 65
|   LYSOZYME < 7.5
|   |   LYSOZYME < 3.5
|   |   |   CHITINASE < 15 : 21.4 (1/0)
|   |   |   CHITINASE >= 15 : 22.2 (1/0)
|   |   LYSOZYME >= 3.5 : 23.2 (1/0)
|   LYSOZYME >= 7.5
|   |   CHITINASE < 15 : 29 (1/0)
|   |   CHITINASE >= 15 : 28.2 (1/0)

Size of the tree : 29
```

**Figure 7.** The trees.RandomTree model.

## 4. Conclusion and Suggestions

For the maximum lipid content from *C. minutissima*, the RSM model predicted process parameters of 20% (v/v) wastewater concentration, 5 U/mL chitinase, and 10 U/mL lysozyme. Under these conditions, the lipid percentage was 30.4% with a very low MSE (0.2414) and a very high $R^2$ value (0.9765). According to the ANOVA results, *P*-values (0.00021134 and 1.5141e$^{-08}$) indicated that wastewater concentration and lysozyme had a positive effect on lipid production at a 95% confidence level, but chitinase had a negative effect. The ANN model, with a minimal experimental setup, can reduce chemical usage and provide low-cost lipid production. Under the conditions of 20% (v/v) wastewater concentration, 5 U/mL chitinase, and 10 U/mL lysozyme, the lipid concentration was 30.3%. The $R^2$ values of the training, validation and test data were 0.99548, 0.96336, and 0.98257, respectively. The effects of parameters were also examined by decision tree ensemble models.

The tRT model had the lowest RMSE of 0.2156 and MAE of 0.1167, with the highest correlation

coefficient of 0.9961. Also, the lipid percentage was 30.3% for the combined RSM-tRT model. The optimal values of wastewater concentration, chitinase, and lysozyme were 20% (v/v), 5 U/mL and 10 U/mL, respectively. Compared to the RSM and ANN models, the tRT model was better at fitting experimental data and generalizing from it. It was also a powerful tool to help make decisions and optimize parameters. The if-then rules extracted from tRT are also given to researchers to test the new parameters in the model. The fuel industry can benefit from the developed models. This work further demonstrates that the combined RSM-tRT model is very close to experimental results and can also be used to obtain biodiesel in a more economical way.

**Acknowledgment**

I would like to thank Van Yuzuncu Yıl University, Department of Environmental Engineering, and Dr. Erdinç Aladağ for the experimental data.

**Conflict of Interest Statement**

The author has no conflicts of interest.

**Statement of Research and Publication Ethics**

The study is complied with research and publication ethics.

**References**

[1] S. Anto, S.S. Mukherjee, R. Muthappa, T. Mathimani, G. Deviram, S.S. Kumar, T.N. Verma, A. Pugazhendhi, "Algae as green energy reserve: Technological outlook on biofuel production," *Chemosphere*, vol. 242, 125079, 2020.

[2] A. Garg, S. Jain, "Process parameter optimization of biodiesel production from algal oil by response surface methodology and artificial neural networks," *Fuel*, vol. 277, 118254, 2020.

[3] G. Srivastava, A.K. Paul, V.V. Goud, "Optimization of non-catalytic transesterification of microalgae oil to biodiesel under supercritical methanol condition," *Energy Convers Manag*, vol. 156, pp. 269–278, 2018.

[4] M. Karimi, "Exergy-based optimization of direct conversion of microalgae biomass to biodiesel," *J Clean Prod*, vol. 141, pp. 50–55, 2017.

[5] S. Nagappan, S. Devendran, P.C. Tsai, H.U. Dahms, V.K. Ponnusamy, "Potential of two-stage cultivation in microalgae biofuel production," *Fuel*, vol. 252, pp. 339–349, 2019.

[6] M. Nayak, G. Dhanarajan, R. Dineshkumar, R. Sen, "Artificial intelligence driven process optimization for cleaner production of biomass with co-valorization of wastewater and flue gas in an algal biorefinery," *J Clean Prod*, vol. 201, pp. 1092–1100, 2018.

[7] S. Chakravarty, N. Mallick, "Optimization of lipid accumulation in an aboriginal green microalga Selenastrum sp. GA66 for biodiesel production," *Biomass and Bioenergy*, vol. 126, pp. 1–13, 2019.

[8] W.B. Kong, S.F. Hua, H. Cao, Y.W. Mu, H. Yang, H. Song, C.G. Xia, "Optimization of mixotrophic medium components for biomass production and biochemical composition biosynthesis by Chlorella vulgaris using response surface methodology," *J Taiwan Inst Chem Eng*, vol. 43, pp. 360–367, 2012.

[9] S. Singh, J.P. Chakraborty, M.K. Mondal, "Optimization of process parameters for torrefaction of Acacia nilotica using response surface methodology and characteristics of torrefied biomass as upgraded fuel," *Energy*, vol. 186, 115865, 2019.

[10] U. Suparmaniam, M.K. Lam, Y. Uemura, J.W. Lim, K.T. Lee, S.H. Shuit, "Insights into the microalgae cultivation technology and harvesting process for biofuel production: A review," *Renew Sustain Energy Rev*, vol. 115, 109361, 2019.

[11] S.O. Ajala, M.L. Alexander, "Multi-objective optimization studies of microalgae dewatering by utilizing bio-based alkali: a case study of response surface methodology (RSM) and genetic algorithm (GA)". *SN Appl Sci*, vol. 2, pp. 1–20, 2020.

[12] A. Kirrolia, N.R. Bishnoi, R. Singh, "Response surface methodology as a decision-making tool for optimization of culture conditions of green microalgae Chlorella spp. for biodiesel production," *Ann Microbiol*, vol. 64, pp. 1133–1147, 2014.

[13] G. Satpati, S.K. Mallick, R. Pal, "An alternative high-throughput staining method for detection of neutral lipids in green microalgae for biodiesel applications," *Biotechnol Bioprocess Eng.* vol. 20, pp. 1044–1055, 2015.

[14] F.J. Chu, T.J. Wan, T.Y. Pai, H.W. Lin, S.H. Liu, C.F. Huang, "Use of magnetic fields and nitrate concentration to optimize the growth and lipid yield of Nannochloropsis oculata," *J Environ Manage*, vol. 253, 109680, 2020.

[15] M.F. Kamaroddin, A. Rahaman, D.J.Gilmour, W.B. Zimmerman, "Optimization and cost estimation of microalgal lipid extraction using ozone-rich microbubbles for biodiesel production," *Biocatal Agric Biotechnol*, vol. 23, 101462, 2020.
[16] Supriyanto, R. Noguchi, T. Ahamed, D.S. Rani, K. Sakurai, M.A. Nasution, D.S. Wibawa, M. Demura, M.M. Watanabe, "Artificial neural networks model for estimating growth of polyculture microalgae in an open raceway pond," *Biosyst Eng*, vol. 177, pp. 122–129.
[17] E. Baldev, D. Mubarakali, K. Saravanakumar, C. Arutselvan, N.S. Alharbi, S.A. Alharbi, V. Sivasubramanian, N. Thajuddin, "Unveiling algal cultivation using raceway ponds for biodiesel production and its quality assessment," *Renew Energy*, vol. 123, pp. 486–498, 2018.
[18] C. Zhang, Y. Zhang, B. Zhuang, X. Zhou, "Strategic enhancement of algal biomass, nutrient uptake and lipid through statistical optimization of nutrient supplementation in coupling Scenedesmus obliquus-like microalgae cultivation and municipal wastewater treatment," *Bioresour Technol*, vol. 171, pp. 71–79, 2014.
[19] P. Polburee, S. Limtong, "Economical lipid production from crude glycerol using Rhodosporidiobolus fluvialis DMKU-RK253 in a two-stage cultivation under non-sterile conditions," *Biomass and Bioenergy*, vol. 138, 105597, 2020.
[20] M. Tourang, M. Baghdadi, A. Torang, S. Sarkhosh, "Optimization of carbohydrate productivity of Spirulina microalgae as a potential feedstock for bioethanol production," *Int J Environ Sci Technol*, vol. 16, pp. 1303-1318, 2017.
[21] C. Huang, H. Wu, R. feng Li, M. hua Zong, "Improving lipid production from bagasse hydrolysate with Trichosporon fermentans by response surface methodology," *N Biotechnol*, vol. 29, pp. 372–378, 2012.
[22] M. Mäkelä, "Experimental design and response surface methodology in energy applications: A tutorial review," *Energy Convers Manag*, vol. 151, pp. 630–640, 2017.
[23] S.K. Yellapu, J. Bezawada, R. Kaur, M. Kuttiraja, R.D. Tyagi, "Detergent assisted lipid extraction from wet yeast biomass for biodiesel: A response surface methodology approach," *Bioresour Technol*, vol. 218, pp. 667–673, 2016.
[24] A. Onay, "Investigation of biomass productivity from Nannochloropsis gaditana via response surface methodology using MATLAB," *Energy Reports*, vol. 6, pp. 44-49, 2020.
[25] S.M. Huang, C.H. Kuo, C.A. Chen, Y.C. Liu, C.J. Shieh, "RSM and ANN modeling-based optimization approach for the development of ultrasound-assisted liposome encapsulation of piceid," *Ultrason Sonochem*, vol. 36, pp. 112–122, 2017.
[26] A.A. Ayoola, F.K. Hymore, C.A. Omonhinmin, O.C. Olawole, O.S.I. Fayomi, D. Babatunde, O. Fagbiele, "Analysis of waste groundnut oil biodiesel production using response surface methodology and artificial neural network," *Chem Data Collect*, vol. 22, 100238, 2019.
[27] M. Mondal, A. Ghosh, K. Gayen, G. Halder, O.N. Tiwari, "Carbon dioxide bio-fixation by Chlorella sp. BTA 9031 towards biomass and lipid production: Optimization using Central Composite Design approach," *J CO2 Util*, vol. 22, pp. 317–329, 2017.
[28] M.A. Alam, J. Wu, J. Xu, Z. Wang, "Enhanced isolation of lipids from microalgal biomass with high water content for biodiesel production," *Bioresour Technol*, vol. 291, 121834, 2019.
[29] N.B. Ishola, A.A. Okeleye, A.S. Osunleke, E. Betiku, "Process modeling and optimization of sorrel biodiesel synthesis using barium hydroxide as a base heterogeneous catalyst: appraisal of response surface methodology, neural network and neuro-fuzzy system," *Neural Comput Appl*, vol. 31, pp. 4929–4943, 2019.
[30] S. González, S. García, J.D. Ser, L. Rokach, F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Inf Fusion*, vol. 64, pp. 205–237, 2020.
[31] S. Das, R. Chakraborty, A. Maitra, "A random forest algorithm for nowcasting of intense precipitation events," *Adv Sp Res*, vol. 60, pp. 1271–1282, 2017.
[32] L. Wang, X. Zhou, X. Zhu, Z. Dong, W. Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data," *Crop J*, vol. 4, pp. 212–219, 2016.
[33] L. Rokach, "Decision forest: Twenty years of research," *Inf Fusion*, vol. 27, pp. 111–125, 2016.
[34] A. Onay, M. Onay, "A Drug Decision Support System for Developing a Successful Drug Candidate Using Machine Learning Techniques," *Current Computer-Aided Drug Design*, vol. 16, pp. 407-419, 2020.
[35] Y. Song, L. Wang, X. Qiang, W. Gu, Z. Ma, G. Wang, "The promising way to treat wastewater by microalgae: Approaches, mechanisms, applications and challenges," *Journal of Water Process Engineering*, vol. 49, 103012, 2022.

[36] W.N.A. Kadir, M.K. Lam, Y. Uemura, J.W. Lim, K.T. Lee, "Harvesting and pre-treatment of microalgae cultivated in wastewater for biodiesel production: A review," *Energy Conversion and Management*, vol. 171, pp. 1416-1429, 2018.
[37] S. I. Khan, I. Hashmi, S.J. Khan, R. Henderson, "Performance and optimization of lab-scale membrane bioreactors for synthetic municipal wastewater," *Desalin Water Treat*, vol. 57, pp. 29193–29200, 2016.
[38] F. Jordi, M. Lees , G.M. Sloane-Stanley, "A simple method for the isolation and purification of total lipids from animal tissues," *J Biol Chem*, vol. 226, pp. 497-509, 1957.
[39] A. Cutler, D.R. Cutler, J.R. Stevens, "Ensemble Machine Learning," *Ensemble Mach Learn*, 2012.
[40] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[41] M. Muthuraj, N. Chandra, B. Palabhanvi, V. Kumar, D. Das, "Process Engineering for High-Cell-Density Cultivation of Lipid Rich Microalgal Biomass of Chlorella sp. FC2 IITG," *Bioenergy Res*, vol. 8, pp. 726–739, 2015.
[42] H.A. Thanaa, S.E.G. Mamdouh, H.E.G. Dina, E.A. Ghada, E.T. Amir, "Improvement of lipid production from an oil-producing filamentous fungus, Penicillium brevicompactum NRC 829, through central composite statistical design," *Ann Microbiol*, vol. 67, pp. 601–613, 2017.
[43] M. Dammak, S.M. Haase, R. Miladi, F.B. Amor, M. Barkallah, D. Gosset, C. Pichon, B. Huchzermeyer, I. Fendri, M. Denis, S. Abdelkafi, "Enhanced lipid and biomass production by a newly isolated and identified marine microalga," *Lipids Health Dis*, vol. 15, pp. 1–13, 2016.
[44] A. Onay, "Optimization of lipid content of Nannochloropsis gaditana via quadratic models using Matlab Simulink," *Energy Reports*, vol. 6, pp. 128–133, 2020.
[45] L. Khaouane, C. Si-Moussa, S. Hanini, O. Benkortbi, "Optimization of culture conditions for the production of pleuromutilin from pleurotus mutilus using a hybrid method based on central composite design, neural network, and particle swarm optimization," *Biotechnol Bioprocess Eng*, vol. 17, pp. 1048–1054, 2012.
[46] M.N.A. Sohedein, W.A.A.Q.I. Wan-Mohtar, Y. Hui-Yin, Z. Ilham, J.S. Chang, S. Supramani, P. Siew-Moi, "Optimisation of biomass and lipid production of a tropical thraustochytrid Aurantiochytrium sp. UMACC-T023 in submerged-liquid fermentation for large-scale biodiesel production," *Biocatal Agric Biotechnol*, vol. 23, 101496, 2020.
[47] S. Chakravarty, N. Mallick, "Optimization of lipid accumulation in an aboriginal green microalga Selenastrum sp. GA66 for biodiesel production," *Biomass and Bioenergy*, vol. 126, pp. 1–13, 2019.
[48] Y. Zhang, X. Kong, Z. Wang, Y. Sun, S. Zhu, L. Li, P. Lv, "Optimization of enzymatic hydrolysis for effective lipid extraction from microalgae Scenedesmus sp.," *Renew Energy*, vol. 125, pp. 1049–1057, 2018.
[49] L. Zhang, B. Chao, X. Zhang, "Modeling and optimization of microbial lipid fermentation from cellulosic ethanol wastewater by Rhodotorula glutinis based on the support vector machine," *Bioresour Technol*, vol. 301, 122781, 2020.
[50] A.M. Nassef, H. Rezk, M.A. Abdelkareem, A. Alaswad, A. Olabi, "Application of fuzzy modelling and Particle Swarm Optimization to enhance lipid extraction from microalgae," *Sustain Energy Technol Assessments*, vol. 35, pp. 73–79, 2019.
[51] J. Ferrell and V. Sarisky-Reed, "National Algal Biofuels Technology Roadmap," U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, *Biomass Program*, 2010.
[52] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, 101805, 2023.