**Predictive Modelling of Biodiesel Production from Waste through Transesterification**

**Authors**

Rutvik Kumara, Rejeti Venkata Srinadha , Neelancherry Remyaa

**Affiliation**

aInstitute of Technology Bhubaneswar, Odisha, India – 752050a

**\*Corresponding Author**

Corresponding author’s email: remya@iitbbs.ac.in

Corresponding author’s Ph. No.: +91-8280549054

# **ABSTRACT**

Biodiesel production and use have been the topic of extensive research due to the ever-increasing emphasis on the creation of eco-friendly and low carbon footprint energy alternatives. Several process factors drive the biodiesel production process, which must be kept at optimal levels to ensure maximum productivity. Because biodiesel productivity and quality are also affected by the numerous raw materials used in transesterification, practical studies are required before making any assumptions about them. However, due to the high number of process parameters and the underlying non-linear relationship between the process parameters and responses, a brute force strategy of carrying out physical trials until the optimal process parameters are attained will fail. In this context, this research employs a machine learning-based prediction approach to quantify the response properties of the biodiesel manufacturing process as a function of process parameters. This paper investigates four powerful machine learning algorithms in depth: linear regression, random forest regression, adaBoost regression and artificial neural network. Both random forest regression and adaBoost regression show good accuracy in prediction modelling of biodiesel yield. However, random forest may be a better strategy for modelling biodiesel production because it gets the lowest error among the studied algorithms. Furthermore, random forest can be deployed more quickly because it was found to be insensitive to the amount of regressors used.

Keywords: biodiesel; machine learning; linear regression; random forest regression; adaBoost regression; artificial neural network

## **1.INTRODUCTION**

Biodiesel is a non-toxic, biodegradable fuel that is aromatic and sulfur-free. The transesterification reaction between waste oil and different alcohols such as methanol or ethanol results in the production of biodiesel. The importance of biodiesel has grown in recent years as the world's total petroleum reserves have decreased and global warming concerns have grown [1]. It has been observed that the use of petroleum is increasing exponentially year after year, which could lead to the depletion of its reserves by 2042 [2]. This perilous situation has piqued the interest of researchers, who are working hard to develop alternatives to petroleum products. Without any major modifications, biodiesel can be considered an alternative to fossil fuels used in diesel engines. Researchers are recently interested in using vegetable oil to produce biodiesel, which is less polluting than petroleum. Biodiesels will be extremely beneficial to the environment and job creation in the local community if they are used effectively for energy generation. The usage of indigenous (for example, Calophyllum inophyllum is native to Asia) or abundant (for example, Brazil and the United States lead in soybean production) biomass sources for biodiesel production encourages farmers to cultivate such crops [3]. Various processes have been used in recent years to produce biodiesel from oils such as soybean oil, canola oil, palm oil, sunflower seed oil, and so on. Transesterification is an important class of organic equilibrium exchange reaction in which one ester is transformed into another by the interchange of the alkoxy moiety [4]. A response surface methodology (RSM) approach has been widely used for optimising process factors to achieve the desired output with fewer experiments. To carry out the number of experiments, assorted designs of experiment (DoE) such as Taguchi experimental design, factorial design, central composite design (CCD) and box-benhken design (BBD) are used. To optimise process factors, various authors used different DoE and statistical techniques using machine learning as it will helps to accurately predict the values without performing the experiment.

**1.1Transesterification process**

Transesterification method is now widely used to extract biodiesel from non-edible oil, waste cooking oil, and animal waste and fat. The oil or fat part of the waste is dissolved with alcohol in this process. Methanol and ethanol are commonly used in this process due to their availability and low cost [6]. Catalyst plays a key role in this reaction to extract biodiesel with a higher yield percentage. Acid, alkali, enzymatic, and base catalysts such as NaOH, KOH, Ca(OH)2, CaCO3, and others are commonly used. According to recent research, CaCO3 is the better catalyst; it has been proven that using CaCO3 as a catalyst, approximately 95% biodiesel can be extracted from animal fat and non-edible oil [7]. Furthermore, new catalysts like rice husk, fishbone, and others can be processed to increase biodiesel yield. Another important fact is that it is a simple and inexpensive process, and net-zero carbon emissions can be achieved if biodiesel is produced using this method [8].

**2. MATERIALS AND METHODS**

Predictive modelling of biodiesel production will require the use of machine learning algorithms and the process involved for experimentation. In this chapter we have covered the method involved for production of biodiesel. It has described about the transesterification process, the different catalyst involved and about the type of alcohol can be used. It also provides the information about the basic machine learning algorithms like linear regression and some advanced algorithms like random forest, adaptive boosting regression and artificial neural network also supplies a detail overview of how this algorithms work. Another method for extracting biodiesel is a supercritical transesterification process. A catalyst is not needed in this process, which only requires high heat and pressure. It is also more environmentally friendly and has a higher yield percentage than the earlier one, but it is more expensive [10-12]. As a result, this method is only used in exceptional circumstances. Table 2.1 gives the sample dataset of process parameters such as temperature, reaction time, yield and catalyst weight from the original dataset which were used for prediction.

Table 2.1. Sample Dataset of Process Parameters and Experimental Yield %

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Catalyst | Methanol/Oil Molar Ratio(Mr) | Catalyst Weight (Wc) | Reaction Temperature (T) | Reaction Time (Tr) | Yield (%) |
|  | 5 | 0.7 | 35 | 45 | 93.3 |
|  | 6 | 1 | 65 | 30 | 95.88 |
|  | 5 | 0.7 | 50 | 30 | 96.62 |
|  | 5 | 1 | 50 | 45 | 94.9 |
| KOH | 6 | 0.4 | 35 | 30 | 96.4 |
|  | 4 | 1 | 65 | 60 | 92.26 |
|  | 6 | 0.4 | 35 | 60 | 96.84 |
|  | 5 | 0.7 | 50 | 45 | 94.22 |
|  | 4 | 1 | 35 | 60 | 91.04 |
|  | 4 | 0.4 | 35 | 60 | 84.14 |
|  | 4 | 0.7 | 50 | 45 | 95.9 |
| NaOH | 6 | 0.4 | 65 | 60 | 98.1 |
|  | 6 | 1 | 35 | 30 | 96.86 |
|  | 6 | 1 | 35 | 60 | 98.72 |
|  | 5 | 0.7 | 50 | 45 | 96.66 |
|  | 6 | 1 | 65 | 60 | 95.5 |
|  | 5 | 0.4 | 50 | 45 | 94.58 |
| KOH | 6 | 0.7 | 50 | 45 | 99.54 |
|  | 5 | 0.7 | 50 | 60 | 95.68 |
|  | 4 | 1 | 35 | 30 | 94.86 |
|  | 6 | 0.4 | 65 | 30 | 98.41 |
| Ca(OH)2 | 4 | 0.4 | 35 | 30 | 85.88 |
|  | 5 | 0.7 | 50 | 45 | 96.86 |
| KOH | 4 | 0.4 | 65 | 30 | 96.48 |
|  | 5 | 0.7 | 65 | 45 | 94.52 |

**2.2Machine Learning Algorithms**

**Linear Regression**

Linear regression may be defined as the statistical model that analyses the linearrelationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable also changes accordingly (increase or decrease) [13].

Mathematically, the relationship can be represented per equation 1:

Y=mX +b (1)

Where, 𝑌 is the dependent variable (to be predicted)

𝑋 is the dependent variable used to make predictions.

𝑚 is the slope of the regression line which stands for the effect 𝑋on 𝑌

𝑏 is a constant, known as the 𝑌-intercept. If 𝑋 = 0, 𝑌 would be equal to 𝑏.

In case the best fit line passes through origin, the equation becomes

Y=bX (2)

Generalized linear regression model for any response (y) and its predictors (𝑥1, 𝑥2, ...,𝑥n), i.e., biodiesel production process parameters in this case can be said as:

Y = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2 + 𝛽n𝑥n (3)

Furthermore, the linear relationship can be positive or negative in nature as explained below:

Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases. Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decreases.

**Random Forest Regression**

The Random Forest is based on applying bagging to decision trees, with one important extension: in addition to sampling the records, the algorithm also samples the variables. To decide how to create a sub partition of a partition A, the algorithm chooses the variable and split point by minimising a criterion such as Gini impurity with random forests, the choice of variable is limited to a random subset of variables at each stage of the algorithm [14]. Compared to the basic tree algorithm, the random forest algorithm adds two more steps: the bagging and the bootstrap sampling of variables at each split:

1. Take a bootstrap (with replacement) subsample from the *records*.
2. For the first split, sample *p* <*P variables* at random without replacement.
3. For each of the sampled variables *Xj*1 , *Xj*2 , ..., *Xjp*, apply the splitting algorithm:

a. For each value *sjk*of *Xjk*:

1. Split the records in partition *A*, with *Xj*(*k*) <*sj*(*k*) as one partition and the remaining records where *Xjk*≥ *sjk*as another partition.
2. Measure the homogeneity of classes within each subpartition of *A*.

b. Select the value of *sjk*that produces maximum within-partition homogeneityof class.

1. Select the variable *Xjk*and the split value *sjk*that produces maximum within partitionhomogeneity.
2. Proceed to the next split and repeat the earlier steps, step 2 onwards.
3. Continue with added splits, following the same procedure until the tree is grown.
4. Go back to step 1, take another bootstrap subsample, and start the process over again.

**Adaptive Boosting Regression**

Adaptive boosting, also known as AdaBoost, is a sequential ensemble technique that develops several weak learners using different training subsets drawn at random from the original training dataset [13-14]. Weights are assigned during each training and are used when learning each hypothesis. The weights are used to compute the hypothesis error on the dataset and act as an indicator of the relative importance of each instance. After each iteration, the weights are recalculated so that instances incorrectly classified by the earlier hypothesis receive higher weights. This allows the algorithm to concentrate on more difficult-to-learn instances.

**Artificial neural network (ANN)**

ANN mimics human brain functionality and takes the same approach as the human brain in recognising patterns, learning, cognition, and making decisions. The human brain is made up of billions of neurons that are all linked together. ANN does the same thing, passing data from an input layer to one or more hidden layers where learning occurs, followed by an output layer that may result in a decision. ANN is commonly used to solve classification and regression problems in supervised learning. The ANN learning system consists of multilayer perceptron’s, backpropagation, radial basis function networks, and resilience. The deep learning method cansignificantly contribute to predicting FW and biofuel generation, which will eventually aid in more efficient system management. Deep learning methods include convolutional neural networks, deep relief networks, Boltzmann machines, and au-to encoders, which could help waste and energy systems [15].

**Predictive Model Performance Evaluation Metrics**

The residue, 𝜀i between the ith original value, “𝑦i” and predicted value, “y0”is calculated as,

𝜀i = 𝑦i – 𝑦0 (4)

The coefficient of determination, 𝑅2 is computed using the 𝑦i, 𝑦0and the mean of the dataset, ym.

𝑅2 = 1 − ∑ (𝑦i− 𝑦0i)2/∑ (𝑦i − 𝑦m)2 (5)

The mean-absolute-error is computed based on the number of the samples, 𝑛as

MAE = ∑ |𝑦i − 𝑦0i|/𝑛 (6)

The mean-squared-error, MSE, is computed as

MSE = ∑ (𝑦i − 𝑦0i)/𝑛 (7)

The root-mean-squared-error, RMSE, is computed as

RMSE = (∑ (𝑦i – y0i)2/𝑛 )0.5 (8)

The maximum error is computed as

Max. Error (𝑦, 𝑦0) = max. (|𝑦i − 𝑦0i|) (9)

The median error is computed as

Med. Error (y, y0) = median (|yi – y01|, ... , |yn– y0n) (10)

**2.3Biodiesel Production Yield Estimation**

The experimental data are from studies in which alkali (KOH) is used as a catalyst. Biodiesel production from waste cooking oil is done with two-step transesterification reaction. The oil was pre-heated for 30 minutes at 110 °C to remove the moisture content. A certain amount of KOH pellets was dissolved in methanol to prepare various percentages of catalyst. The preheated oil was cooled to specific temperatures (30, 50, and 65 °C) before being mixed with a potassium methoxide (CH3KO) solution. In a 250 mL three-neck glass reactor, the CH3KO solution was mixed with the pre-heated oil and fixed over a magnetic stirrer with constant stirring at the above specific temperature. The mixture was then held at three different reaction times (30, 45 and 60 min). The above procedure was repeated three times to ensure the removal of contaminants from the biodiesel. Following that, the biodiesel was heated for 1 hour at 100 °C to allow residual water to evaporate. The fraction with 0.1 N sulphuric acid was used to quantify the biodiesel titration of biodiesel. Standard procedures were used for the physicochemical characterization of oil and biodiesel. Oil was analysed to figure out the fatty acid constituents. For the design of experiments in this case, central composite design (CCD) was considered. CCD can be performed in two ways: face-centred central composite design (FCCD) and rotatable central composite design (RCCD). In this case, FCCD is used to decide the effect of various process parameters on the largest biodiesel yield. Two levels and four factors with five different centred point values were considered for the experiment's design. The discussed process parameters chosen for the experiments are reaction time, methanol/oil molar ratio (M), reaction temperature, and catalyst weight percentage.

## **4. RESULTS AND DISCUSSION**

Evaluating the performance of different experimentation with model is must to get the best one for prediction. This chapter includes the information about how much these parameters are correlated with each other and the performance of different machine learning algorithms using this process parameters for predicting the biodiesel yield. This chapter also includes the best process parameters to achieve the largest biodiesel yield.

**4.1Effect of Process Parameters**

Figure 4.1 and depict the training data in the form of scatter plots. Because the experimental design was based on a central composite design, the process parameters for biodiesel production are seen to be distributed primarily across three levels in each case. The scatter plot of yield versus methanol/oil molar ratio (M) shows that the yield increases as M increases. However, the effect of various levels of process parameters is inconclusive in the other three scatter plots of yield versus process parameters. The yield histogram shows that the median of all experimentally calculated yields is around 96%.

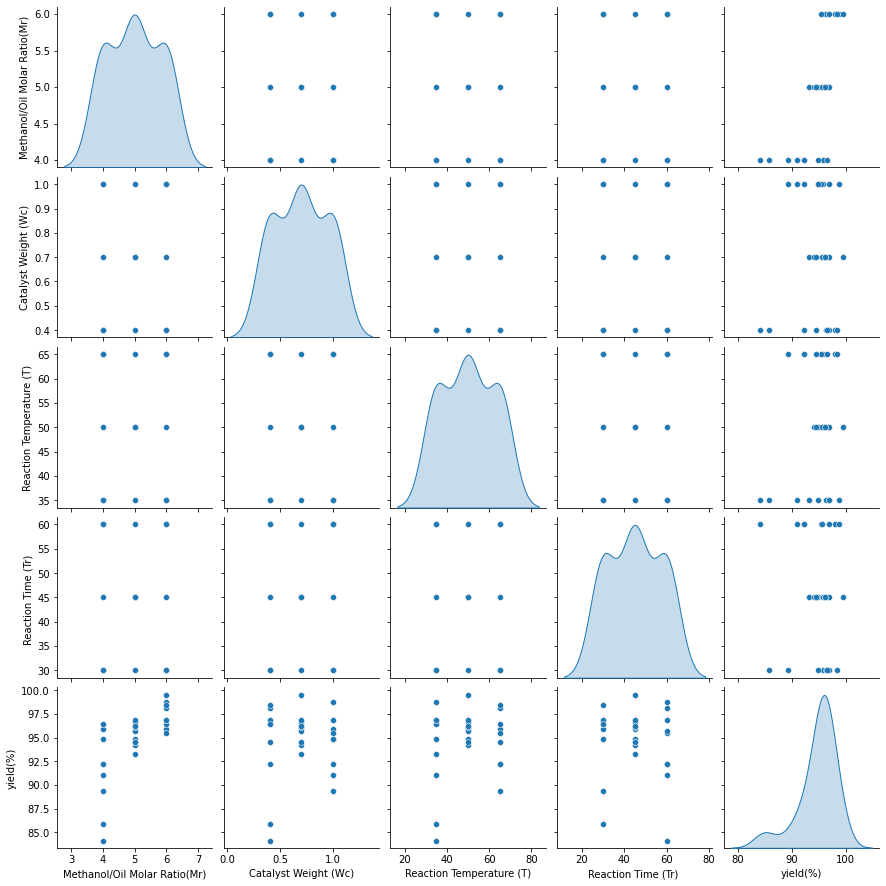


Fig 4.1 Scatter plot of the various process parameters and yield.

Figure 4.2 investigates the relationship between various process parameters and yield. The process parameters have a negligible correlation, confirming the lack of multicollinearity. The methanol/oil molar ratio (M) has a moderately strong positive correlation with yield (0.66). The yield has a negligible relationship with catalyst weight and reaction time. This shows that, on average, low temperatures will not result in a high biodiesel yield. As a result of the analysis of the experimental data on biodiesel yield, as shown in Figures 4.1 and 4.2, a brute force approach to decide the best process parameter combination to maximise yield will be extremely costly. This proves the importance of advanced machine learning algorithms in developing predictive models to quantify yield as a function of process parameters.

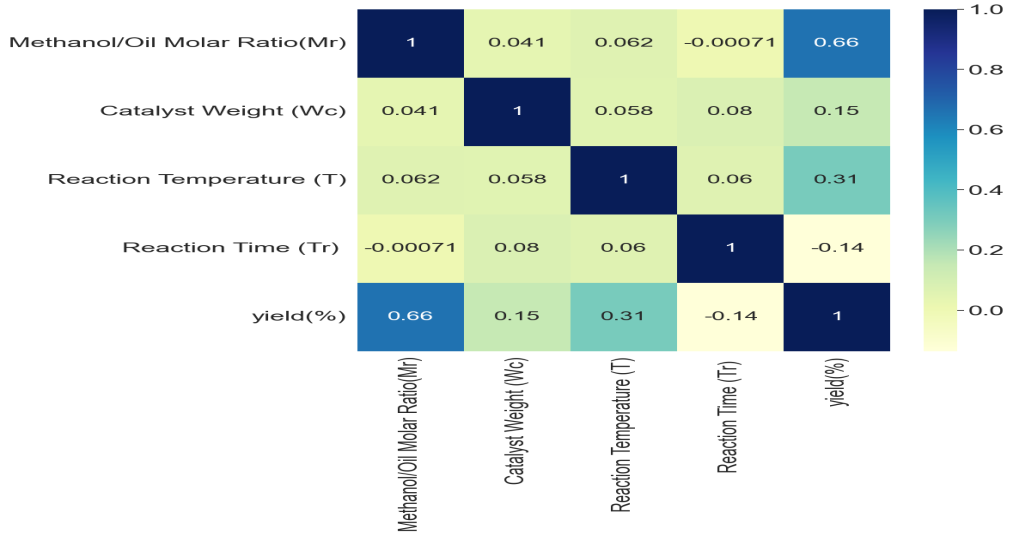


Fig 4.2 Correlation between the various process parameters and yield.

**4.2Linear Regression Predictive Model**

Based on the sample training data, a linear regression predictive model of the yield % was developed. An empirical relation describing the yield % as a function of the methanol/oil molar ratio (𝑀r), catalyst weight (𝑊c), temperature (𝑇) and reaction time (𝑇r) is developed as shown in Equation.

𝑌𝑖𝑒𝑙𝑑 % = 𝛽0 + 𝛽1𝑀r + 𝛽2Wc+ β3T+ β4Tr (11)

Figure 4.3 shows the predictive performance of the linear regression predictive model. The actual yield % is plotted against linear regression predictive model-based predicted yield (%). The best line fit for the data points is generated and the 𝑅2 is calculated which denotes the goodness of fit of the predictive model on the training data. It is seen that the predictive power of the linear regression predictive model is moderate, with a 𝑅2of about 0.52. It is seen that the predicted data points are quite far from the actual data points leads to higher value of residual and make the model less reliable.

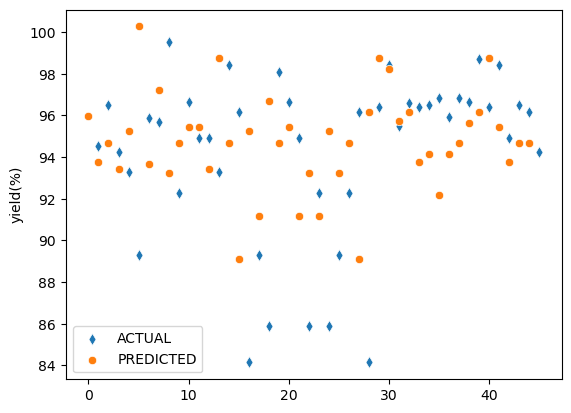
****

Fig 4.3 Predictive performance of the linear regression predictive model

**4.3 Random Forest Regression Predictive Model**

Using the training data, random forest regression predictive models are developed to quantify the yield % as a function of methanol/oil molar ratio (𝑀r), catalyst weight (𝑊c), temperature (𝑇) and reaction time (𝑇r). Figure 4.4 shows the predictive pattern of the selected random forest predictive model plotted between the actual yield% and the predicted yield%. A high 𝑅2 of approximately 98.6 % is achieved with random forest predictive model. Most of the data points are seen to lie on the identity line, showing the high correlation between actual and predicted results. The error in random forest predictive model is much lower than the linear regression predictive model which makes it more reliable than any other model.

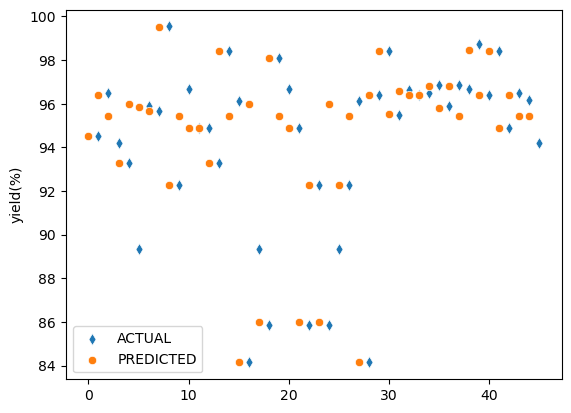


Fig 4.4 Predictive performance of the random forest regression predictive model

**4.4 AdaBoost Regression Predictive Model**

AdaBoost regression predictive model is developed using data. Figure 4.5 shows that the predicted data points are near to the actual data points. The 𝑅2 value of the AdaBoost regression predictive model is approximately 91.2%, showing that the AdaBoost regression predictive model can successfully model 91.2% variance in the training data making it more reliable than linear regression model. The plot in Figure 4.5 shows a balanced prediction for most of the cases, without much under prediction or over prediction.

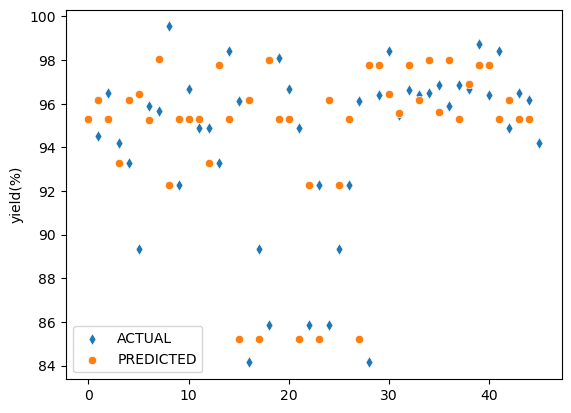


Fig 4.5 Predictive performance of the AdaBoost regression predictive model

**4.5 Artificial Neural Network Model**

An artificial neural network model is developed using the 5 neuron, normal kernel optimizer and relu activation function with 400 epochs. A relation is developed between the yield % as output and the input as the methanol/oil molar ratio (𝑀r), catalyst weight (𝑊c), temperature (𝑇) and reaction time (𝑇r). Figure 4.6 shows the actual and predicted values of yield (%). It seems that error is quite high as the actual and predicted points are too far and the R2 value is 45% making it the bad model among all others.

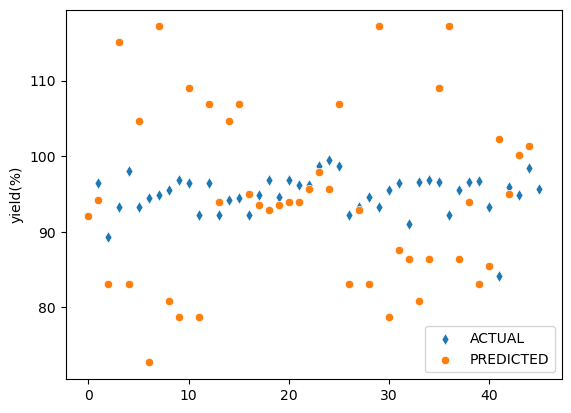


Fig4.6 Predictive performance of the ANN regression predictive model

**4.6 Comparison of Various ML Predictive Models**

The comparison of the R2 for the four predictive models in Figure 4.7(a) reveals that the Random Forest regression predictive model has the best estimation power. The MSE of the Random Forest regression predictive model is almost zero, while for linear regression predictive model, it is around 6. Comparison of the predictive models based on all the error metrics, as shown in Figure 4.7(b), 4.7(c) and 4.7(c)writes down that random forest regression predictive model is comprehensively superior as compared to adaboost forest regression predictive model, ANN regression predictive model and linear regression predictive model.

Fig 4.7 Performance of the models (a) R2; (b) MAE; (c) RMSE; (d) MSE.

**4.7 Best value of features for maximum yield (%)**

Random forest regression model is used to get the best values of the methanol/oil molar ratio (𝑀r), catalyst weight (𝑊c), temperature (𝑇) and reaction time (𝑇r) for maximum yield (%). Table 4.1 shows the 4-maximum value of yield (%) obtained from the predictive model on various values of input parameters. The maximum value of yield (%) in dataset was in the range of 95% to 96% while from the predictive model the highest value of yield (%) is around 99.5%. It has been seen that by keeping the methanol/oil molar ratio (𝑀r) between 6 and 8, the catalyst weight (𝑊c) in the range 0.7 to 0.8, temperature (𝑇) between 45°C and 55°C, reaction time (𝑇r) between 40 min and 50 min and using KOH catalyst will get the maximum values of yield (%).

Table 4.1 Best Predicted Values of Input Parameters using random forest

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Catalyst |  | Methanol/Oil Molar Ratio (Mr) | Catalyst Weight (Wc) | Reaction Temperature (T) | Reaction Time (Tr) | Yield(%) |
|  |  | **7** | 0.8 | 50 | 50 | 99.51 |
| KOH |  | **7** | 0.8 | 60 | 40 | 98.84 |
|  |  | **9** | 0.4 | 35 | 50 | 97.13 |
|  |  | **6** | 0.4 | 25 | 60 | 96.86 |

## **5. CONCLUSION**

Building predictive models for the quantification of biodiesel yield based on the level of the process parameters is a realistic goal with a tremendous impact on sustainable development. In this paper, the biodiesel yield is modelled as a function of methanol/oil molar ratio (𝑀r), catalyst weight (𝑊c), reaction temperature (𝑇)and reaction time (𝑇r). A comprehensive comparative analysis is carried out to learn the utility of four machine learning techniques (linear regression, random forest regression, AdaBoost regression and artificial neural network) for developing predictive models in biodiesel production. A wide range of accuracy and error metrics is used to quantify the efficacy of the machine learning algorithms. Based on the analysis it is seen that the linear regression approach and artificial neural network are only able to achieve moderate accuracy with R2 value of 0.51 and 0.45 respectively, while both random forest regression and AdaBoost regression show extremely high accuracy in predictive modelling of the biodiesel yield. Using random forest for best process parameters it has been seen that by using KOH catalyst with methanol/oil molarratio (𝑀r) of 7, catalyst weight (𝑊c) of 0.8, reaction temperature (𝑇) of 50°C and reaction time (𝑇r) of 50 will give the yield of 99.51%. Thus, Random Forest regression may be preferred for the predictive modelling of biodiesel yield. This would lead to a significant saving in time and effort in finding the Optimal process parameters to increase the yield.

**REFERENCES**

**Journal Paper**

1. Bisheswar Karmakar, Sumit H.Dhawane, Gopinath Halder (2018), “Optimization of biodiesel production from castor oil by Taguchi design” ScienceDirect S2213343718301957
2. Monzur Alam Imteaz, A.B.M. Sharif Hossain, Maryam Bayatvarkeshi (2022), “A mathematical modelling framework for quantifying production of biofuel from waste banana”
3. Min-Yuan Cheng, Doddy Prayogo(2012), “Optimizing Biodiesel Production from Rice Bran Using Artificial Intelligence Approaches”
4. Rouf Ahmad Dar, Mudasir Yaqoob, Manisha Parmar, Urmila Gupta Phutela(2019), “Biofuels from Food Processing Wastes”
5. Navdish Arora(2021), “Biofuels From Food Waste – A Sustainable Approach”
6. Kambiz Tahvildari, Yasaman Naghavi Anaraki, Reza Fazaeli, Sogol Mirpanji, Elham Delrish(2015), “The study of CaO and MgO heterogenic nano-catalyst coupling on transesterification reaction efficacy in the production of biodiesel from recycled cooking oil”
7. Hemal Chowdhury, Tamal Chowdhury, Pranta Barua(2021), “Biofuel production from food waste biomass and application of machine learning for process management”
8. Leung, D, Guo, Y(2006), “Transesterification of neat and used frying oil: Optimization for biodiesel production”
9. Lertsathapornsuk, Pairintra, Aryusuk, Krisnangkura(2008), “Microwave assisted in continuous biodiesel production fromwaste frying palm oil and its performance in a 100 kW diesel generator”
10. Karmakar, B.; Dhawane, S.H.; Halder, G(2018), “Optimization of biodiesel production from castor oil”
11. Kumar, R.; Tiwari, P.; Garg, S.(2013), “Alkali transesterification of linseed oil for biodiesel production”
12. Ahmad, T.; Danish, M.; Kale, P.; Geremew, B.; Adeloju, S.B.; Nizami, M.; Ayoub, M. (2020), “Conversion of flaxseed oil into biodieselusing KOH catalyst: Optimization and characterization dataset”.
13. Hazra, A.; Gogtay, N(2016), “Correlation and linear regression”.
14. Seo, D.K.; Kim, Y.H.; Eo, Y.D.; Park, W.Y.; Park, H.C.(2017),“Generation of Radiometric, Phenological Normalized Image Based onRandom Forest Regression for Change Detection”.
15. Ortega, M.F.; Donoso, D.; Bousbaa, H.; Bolonio, D.; Ballesteros, R.; García-Martínez, M.-J.; Lapuerta, M.; Canoira, L.(2020), “Optimized Production by artificial neural network”.