



Enhancing the dilute acid hydrolysis process using a machine learning approach: investigation of different biomass feedstocks influences glucose and ethanol yields

Nithianantharaj Vinitha¹ · Jaikumar Vasudevan¹ · K. P. Gopinath² · J. Arun³ · S. Madhu⁴ · S. Naveen⁴

Received: 22 February 2024 / Revised: 28 April 2024 / Accepted: 29 April 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

The study focuses on enhancing the efficiency of bioethanol production as a means to reduce reliance on crude oil and mitigate environmental pollution. A key aspect involves the optimization of bioethanol production through the application of an artificial intelligence approach. The Artificial Intelligence Decision-Making System (AIDMS) algorithm was developed using a machine learning algorithm, utilizing datasets derived from experimental results and published research. In the optimization process, a Pearson correlation coefficient matrix was established for 250 training datasets, revealing positive and negative correlation coefficient values. These values underscore the significance of each parameter in ethanol production. Various biomass feedstocks, including cotton stalk, wheat straw, olive tree, potato peel waste, rice straw, and sugarcane bagasse, were selected for validation of the AIDMS algorithm. The validation process compared experimental results with predictions made by the AIDMS, demonstrating a commendable 94% accuracy. The weighted rank order aggregate analysis revealed that cellulose (%), s-temp (°C), acid conc. (%), lignin (%), s-time (min), and hemicellulose (%) show the importance of parameters in obtaining glucose yield. Similarly for ethanol yield, cellulose (%), f-temp (°C), f-time (h), lignin (%), and hemicellulose (%) show the order of rank and its importance. The artificial intelligence-based optimization method is suitable for bioethanol production.

Keywords Machine learning · Optimization · Biomass · Dilute acid hydrolysis

1 Introduction

In the modern world, non-renewable energy sources like fossil fuels are used to meet the world's energy needs [1]. Fossil fuels do, however, have some significant drawbacks, such as

an adverse impact on the environment [2, 3]. Fossil fuels are becoming less abundant due to rising industrialization and population, which is one of the primary contributors to climate change and greenhouse gas emissions. In light of this, it is essential to investigate and optimize the alternate energy sources including biodiesel, bioethanol, and hydrogen [4–6]. The cellulose consisting of polysaccharides and hemicellulose plays an important role in bioethanol production [7, 8]. The conversion of cellulose into glucose and ethanol requires an efficient hydrolysis process [9]. Hence, bioethanol production requires an effective pre-treatment [10]. The change in the structural components of the lignocellulose by pre-treatment is to remove hemicellulose and lignin. Several pre-treatment techniques such as enzymatic hydrolysis, acid hydrolysis, and alkali hydrolysis were used [11]. For the removal of lignin, it requires high chemical concentration and temperature. The development of metabolic inhibitory components is required, which is considered a drawback.

Hydrolysis of lignocellulose for bioethanol production can be carried out in enzymatic and acid hydrolysis [12–14].

✉ S. Madhu
mathumarine@gmail.com
Jaikumar Vasudevan
jaikumaphd@gmail.com

¹ Sri Sivasubramanian Nadar College of Engineering, Chennai 603110, Tamil Nadu, India

² Mohamed Sathak Engineering College, Keelakarai 623806, Tamil Nadu, India

³ Centre for Waste Management-International Research Centre, Sathyabama Institute of Science and Technology, Chennai 600119, Tamil Nadu, India

⁴ Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India

Pre-treatment of lignocellulosic materials can be done using acid hydrolysis. Conversion of hemicellulose into glucose with higher reactions and less cost by using sulfuric acid, hydrochloric acid, and citric acid. At moderate temperatures of acid hydrolysis, saccharification suffers because of the decomposition of glucose [15]. At high temperatures and low pH of acid hydrolysis, it is necessary to utilize high-cost equipment. So, neutralization of acid hydrolysis before the fermentation process is important [16]. Intelligent computerized models connect the gap between research approaches and the latest methods by imparting the ability of optimization to advanced experimental results [17, 18]. Compared to hard computerized models, intelligent user models offer high accuracy in predicting results [17, 19]. Thus, a user-friendly computerized model plays an important role in the optimization of bioethanol production to predict the glucose and ethanol yields. A machine learning algorithm plays a better role when comparing another optimization algorithm [20, 21]. Firstly, food crops, lignocellulosic biomass, and waste are feedstocks of the first and second generations. Microalgae constitute the third and fourth generation of feedstock. In 3rd and 4th generation feedstock, this AIDMS algorithm can be implemented in anaerobic digestion process parameter optimization to obtain biogas yield.

Machine learning is a favorable tool for modeling and optimizing bioethanol production. It can process and predict without prior knowledge, making the method useful for bioethanol production [22, 23]. AI models used in various processes have achieved better results [24]. With the development of a machine learning algorithm, good results were obtained from the fermentation of glucose to estimate the bioethanol production [25]. A machine learning algorithm makes the system analyze the given inputs and change the process conditions to improve efficiency [26]. Machine learning algorithms can be applied in several applications, such as the chemical industry, petroleum industry, manufacturing industry, and climatic conditions [17, 27, 28].

In this work, an Artificial Intelligence Decision-Making System was proposed to obtain the optimal bioethanol yield from the given biomass characteristics and process conditions. In Python programming, the simple linear regression (SLR) and the weighted rank order aggregate (WROA) function were developed. The SLR function categorizes the training dataset consisting of biomass, and the categorized results were stored in the cloud repository. The WROA function allocates weighted rank to the biomass characteristics and process conditions based on the importance of distance weight. The judgment block (JB) was created to admit data from the cloud repository and acts as a port with SLR and WROA functions. Using AIDMS in the optimization of biomass characteristics, saccharification conditions and fermentation conditions. Regression analysis was reported in other research articles. In this study, a ranking order was given to cellulose,

hemicellulose, lignin, and saccharification and fermentation conditions based on the importance of bioethanol production. This AIDMS algorithm considers the importance according to the rank order aggregate and makes brilliant decisions.

2 Materials and methods

This study was by relevant institutional, national, and international guidelines and legislation.

2.1 Collection and preparation of biomass

The biomass feedstock such as cotton stalk, wheat stalk, rice straw, banana plant waste, corn cob, olive tree, potato peel waste, and coconut shells used in this study was collected from a farm around SSN College of Engineering, Kalavakkam, Chennai, Tamil Nadu, India (12.7517° N, 80.2033° E). The biomass feedstocks collected were washed and dried in a hot air oven at 40 °C until the steady weight was accomplished and processed to a 70-mesh size and stored in plastic bags at a dry place until further use. The biomass characteristics and saccharification and fermentation conditions act as the input. As per the standards of the American Society for Testing and Materials (ASTM D 2974-07a), the biomass feedstock was tested for their contents of moisture (%) and ash (%), and the biomass characteristics such as the cellulose (%), hemicellulose (%), and lignin (%) were analyzed using the Neutral Detergent Fiber (NDF) method [29, 30].

2.2 Dilute acid hydrolysis

The upper and lower limits of various operational parameters such as dilute acid concentration, temperature, and retention time were fixed based on previous research works. For acid hydrolysis experiments, a constant solid-to-liquid ratio of 5% (wt/vol) was maintained throughout, and the acid concentration, temperature, and retention time were varied from 0.89 to 5% (vol/vol), 70–120 °C and 20–120 min, respectively [31]. All experiments were carried out in screw-capped Pyrex bottles, and deionized water was added to the mixture to adjust the solid:liquid ratio. Further, the solid and liquid parts were separated by the Buchner funnel after cooling down to room temperature. The reducing sugar concentrations of the hydrolysate (liquid fraction) were determined by Dinitro salicylic acid (DNS) assay with modification [32] method using a UV–Vis spectrophotometer (UV1800, Shimadzu, Japan) at 540 nm [33]. Finally, the hydrolysate required for fermentation for the production of ethanol was prepared by carrying out the above-mentioned acid hydrolysis process.

2.3 Fermentation

The hydrolysate obtained from the method described in Section 2.2 was made to undergo a fermentation process, and ethanol production potential was analyzed. The fermentation process was carried out in Erlenmeyer flasks of 150 mL capacity, having a suitable media of 100 mL at 150 rev/min on a shaker. The operating temperature and retention time varied from 25–35 °C and 24–55 h, respectively. A 4% (v/v) of *Saccharomyces cerevisiae* was used as inoculum (10 ml) with a fermentation medium containing 3 g/L yeast extract, 5 g/L peptone, 2 g/L KH_2PO_4 , 0.2 g/L MgCl_2 , and 0.2 g/L CaCl_2 . Periodically, samples were withdrawn to estimate the amount of ethanol produced by the spectrophotometric potassium dichromate method [34].

2.4 Dataset pre-processing

Data pre-processing steps are applied to the training dataset.

2.4.1 Cleaning of biomass dataset

This involves identifying and correcting errors in the dataset by predicting the missing values using the oversampling method.

2.4.2 Normalization of dataset

It splits the data into training and test datasets. By standardizing the range of independent variables or features within a dataset, data normalization enhances the consistency and comparability of the prediction model and produces more consistent and reliable output.

2.4.3 Dataset feature scaling

Ensuring that every characteristic receives equal consideration during the learning process is made possible by scaling features. Without scaling, learning could be dominated by features at a larger scale, leading to skewed results. Scaling ensures that each feature contributes equally to model predictions while also eliminating this bias. From these three data pre-processing techniques, the accuracy of the results is improved.

2.5 Artificial intelligence algorithm

This algorithm architecture includes a supervised machine learning approach that may be applied to regression and classification problems called support vector machine (SVM). The primary goal of SVM is to identify the optimal boundary, also known as a hyperplane, for classifying the data. It can handle smaller datasets and may effectively

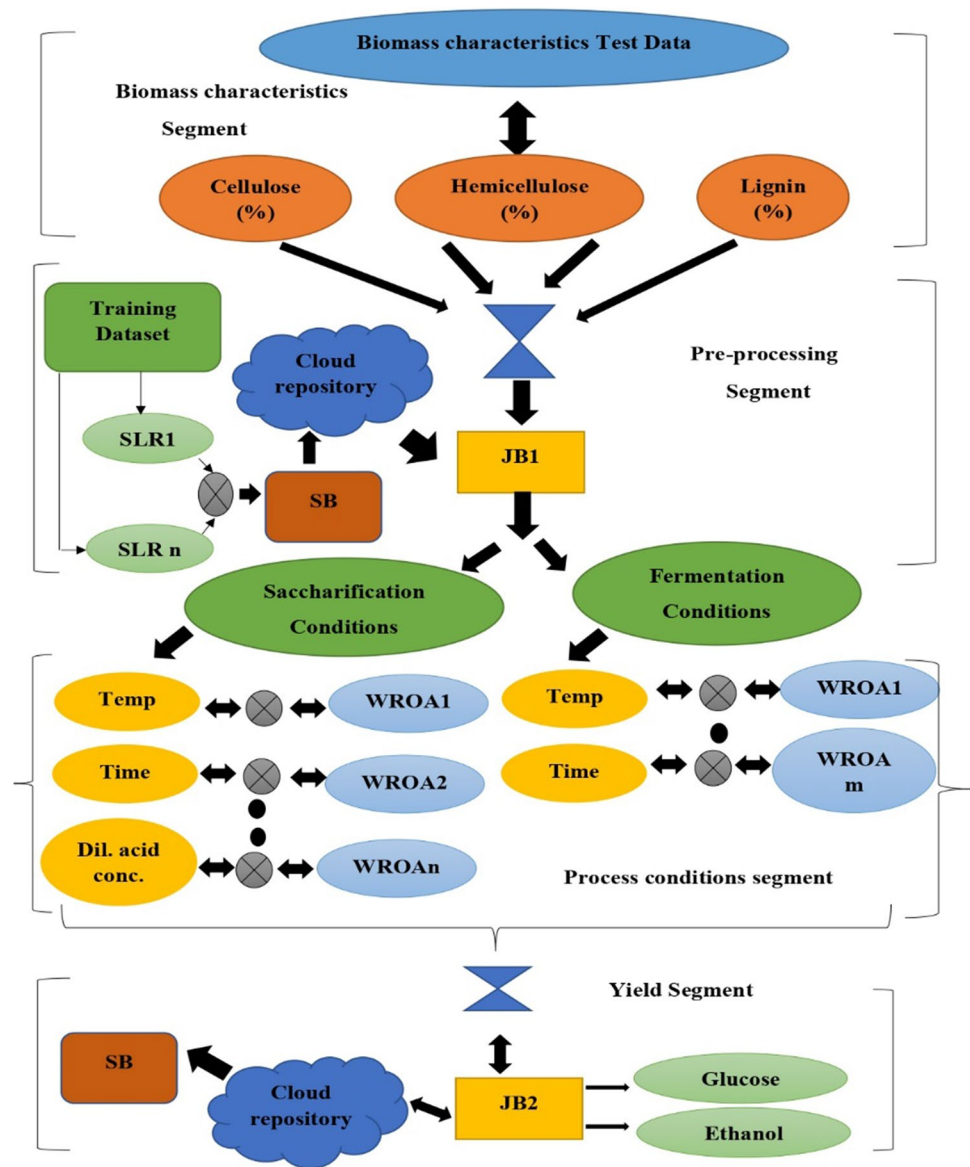
approximate decision boundaries even in the absence of a universal approximation theorem resulting in good accuracy [35]. The main reason for not including ANN in this study because it requires a larger number of labeled datasets for training resulting in less accuracy when compared to SVM. So SVM is an effective and simple tool in the optimization of the dilute acid hydrolysis process. The process parameters like temperature and time symbolize the change in time and temperature. For optimization of dilute acid hydrolysis process parameters to obtain the glucose and ethanol yields, an Artificial Intelligence Decision-Making System was developed. The data containing biomass characteristics, saccharification conditions, and fermentation conditions were collected from various research articles. This study consists of four segments as shown in (Fig. 1), namely, the biomass characteristics segment, the pre-processing segment, the process conditions segment, and the yield segment.

For tuning hyperparameters, the Bayesian optimization model was used because it improves efficiency by evaluating all possible hyperparameter combinations by building the probability model of the objective function [36]. This model also reduces the required evaluations by directing the search toward promising hyperparameters. It is more efficient when compared to other hyperparameter tuning models like random and grid search [37]. This model's main advantage is determining the next dataset to evaluate and automatic tuning [38]. The hyperparameters like cellulose (%), hemicellulose (%), lignin (%), acid concentration (%), saccharification–time (min), saccharification–temperature (°C), glucose (g/L), fermentation–time (h), fermentation–temperature (°C), and ethanol (g/L) were tuned using this model. It improves the AIDMS algorithm performance and interpretability and reduces the overfitting problems.

Biomass characteristics were compared with the training dataset with the help of an SLR function classifier and stored in the storage block called a cloud repository. The obtained results from the SLR act as the input to the bioethanol production module via JB-1. WROA for process conditions is created and acts as the input to the JB-2. The optimum predicted glucose and ethanol yields were stored in the storage block (i.e.) cloud repository. The training dataset consists of biomass feedstock such as rice straw, rice husk, sugarcane bagasse, olive tree, and cotton stalk which was obtained from the tensor flow (".csv") file format and acts as the input to SLR function developed in Python [39]. Sqlite3 is used for operating the database. Pandas' library package is used to manipulate the dataset and acts as the analyzing tool [40]. The nearest neighbor approach uses the function to produce the missing data.

The SLR function is considered for four biomass characteristics such as cellulose (%), hemicellulose (%), lignin (%), and three process conditions for saccharification such as dilute acid concentration (%), temperature (T in °C), and

Fig. 1 Artificial Intelligence Decision-Making System (AIDMS) used in the study



time (t in min), and two process conditions for fermentation such as temperature (T in $^{\circ}\text{C}$) and time (t in h). This SLR function generates an integer subset. This subset consists of cellulose (%) as $C(\%) = (C_1, C_2, \dots, C_i)$ where C_i is considered as the cellulose integer characteristics variable. Similarly, the same integer set is generated for hemicellulose (%) as $H(\%) = (H_1, H_2, \dots, H_i)$ and lignin (%) as $L(\%) = (L_1, L_2, \dots, L_i)$. The acid hydrolysis process conditions such as saccharification and fermentation also have the integer condition variable. Saccharification process conditions also have integer variables such as Acid Conc., S-T, and S-t, (S- conc. as acid concentration, S-T as saccharification temperature, S-t as saccharification time), and for fermentation conditions has F-T and F-t, (F-T as fermentation temperature, F-t as fermentation time). JB deals with the test dataset. WROA allocates rank order to the biomass characteristics

and process conditions according to their importance in the production of bioethanol. The algorithm generated 250 data points, which are detailed in Table A1 (Annexure), designated for training purposes, alongside an additional 20 data points allocated for testing. Euclidean distance algorithm was used in this WROA to obtain the weighted rank for biomass characteristics and process conditions, and the results were validated with the training dataset to obtain the output [41]. The Artificial Intelligence Decision-Making System algorithm analyses every parameter in the dataset to predict the glucose and ethanol yields. The predicted data using the AIDMS algorithm is stored in the storage block of the cloud repository as the additional training dataset. During the optimization procedure, N datasets are predicted and stored in the cloud. This proves that this AIDMS algorithm improves the quality of the dataset.

3 Results and discussion

3.1 Influence of biomass characteristics and process conditions on glucose and ethanol yields

Biomass characteristics such as cellulose (%), hemicellulose (%), and lignin (%) were considered the most efficient parameters in glucose and ethanol yields. Figures 2 and 3 represent the input statistical data analysis used in the study. Five responses such as glucose and ethanol yields were compared with the process conditions such as S-conc., S-T, S- t, F-T and F-t. Because of its higher glucose yield, dilute acid hydrolysis is used in this AIDMS algorithm. Mineral acid like H_2SO_4 can be used in the hydrolysis of cellulose to obtain an ethanol yield. H_2SO_4 is 96% by weight in an aqueous solution [42]. It also consists of a high concentration of H^+ and it is considered a strong catalyst because of its lower water content and would drive the equilibrium towards the reactants [43]. The benefit of ethanol production from cellulose is the inexhaustibility and advantage of cellulosic biomass [44]. Acid hydrolysis of cellulose using sulfuric acid has two steps. In the first step, the acid breaks the β -1,4-glycosidic bond, and the second step is the conversion of cellulose to glucose [45–47]. Acid hydrolysis of hemicellulose consists of two steps. In the first step, 0.7% of H_2SO_4 was used at 190 °C

to hydrolyze the hemicellulose. In the second step, 0.4% of H_2SO_4 was used at 215 °C to obtain the cellulose fraction. Finally, the hydrolyzates are removed before fermentation conditions [48, 49].

During acid hydrolysis, lignin contains more acid, which leads to difficulty in using it as a feedstock. So, lignin is left as residue [50]. Many previous research articles used in this study support this analysis. Cotton stalk biomass had cellulose (40.1%), hemicellulose (13.6%), and lignin (29.4%) content and upon bioethanol production, resulted in (29.4 g/L) of glucose yield [51]. The biomass of *Curcuma longa* comprised cellulose (50%), hemicellulose (11.2%), and lignin (12%) content resulting in (31 g/L) of glucose yields [52]. Cellulose is the important element that forms glucose units with polymerization available in glucose and ethanol yields. Hence higher amount of cellulose content in biomass characteristics is favorable. During the dilute acid hydrolysis-saccharification and fermentation processes, the cellulose (%) is directly proportional to the glucose and ethanol yields, whereas the hemicellulose (%) and lignin (%) are inversely proportional to the glucose and ethanol yields [53]. In the saccharification process, the glucose level attains its maximum at low temperature (70–84 °C). If there is an increase in temperature with increases in acid concentration, the glucose yield decreases are reported in the literature [42]. At a time period of 20–32 min, maximum

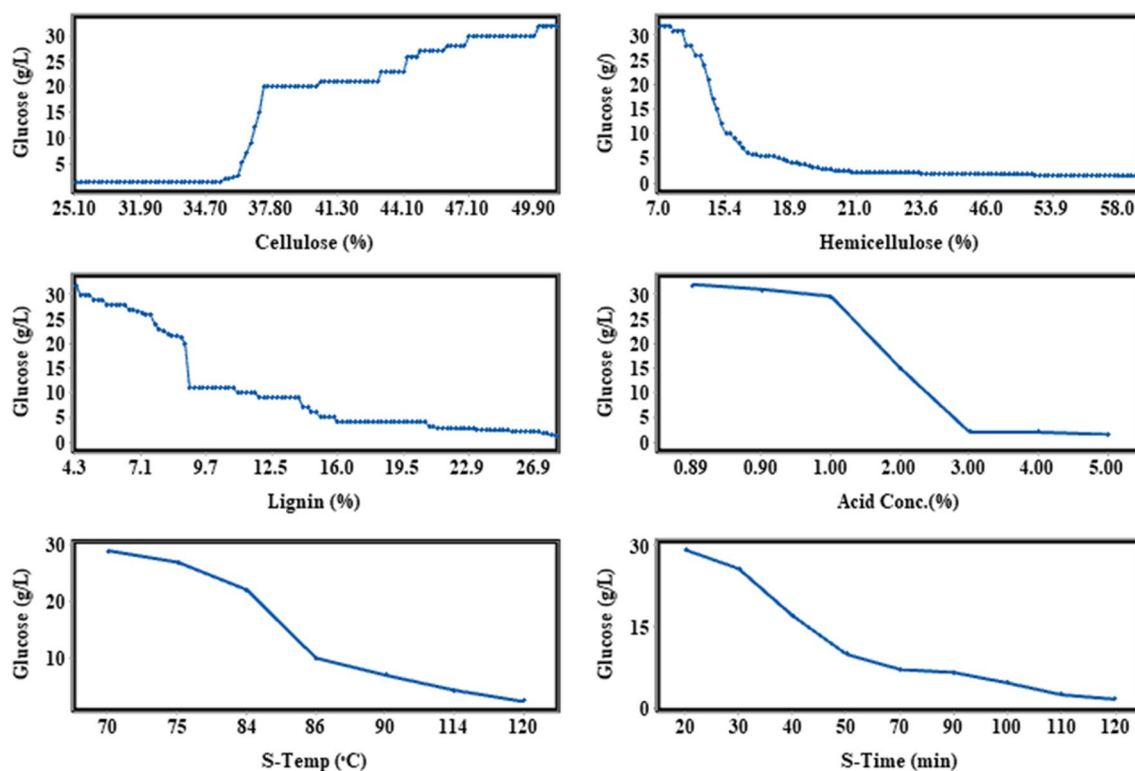


Fig. 2 Influence of biomass characteristics and saccharification conditions on glucose yield (S-saccharification)

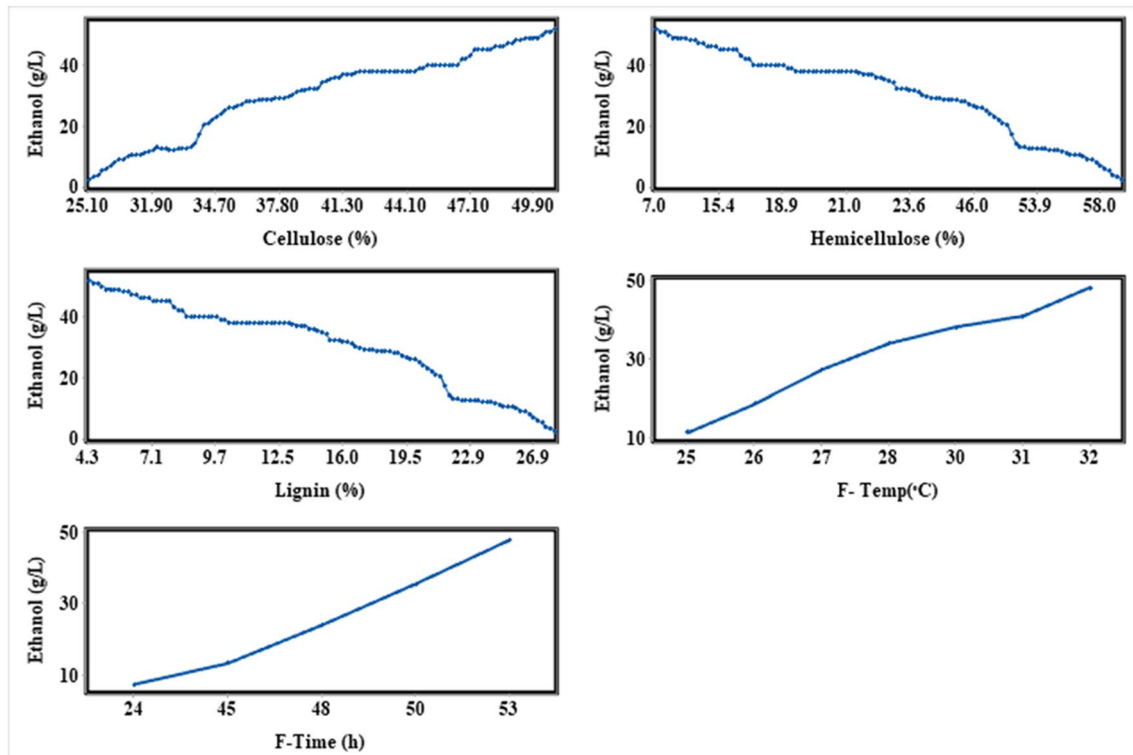


Fig. 3 Influence of biomass characteristics and fermentation conditions on ethanol yield (F-fermentation)

glucose yield was attained. In this analysis, the temperature $< 120\text{ }^{\circ}\text{C}$ and time $< 120\text{ min}$ were taken to reduce the Hydroxymethyl furfurals (HMF) production. Below $120\text{ }^{\circ}\text{C}$ temperature and time of 120 min, the production of HMF is less than 5% reported in the literature [31]. During the fermentation analysis, temperature and time are directly correlated to ethanol yield.

3.2 Correlation coefficient parameters

Analyzing the correlation coefficient values between the biomass characteristics and process conditions and their influence on bioethanol production is important in determining the rank order by using the WROA function. Pearson matrix determines the correlation coefficient values by measuring the strength of the relationship between every variable present in the biomass dataset. It uses the value of one parameter to predict the value of the other parameter. Correlation coefficients are a quantitative assessment that measures both the direction and the strength of this tendency to vary together. It can be used to test whether the relationship between two variables is significant. It has +ve, -ve, and no correlation values. The positive correlation values show the importance of biomass characteristics and dilute acid saccharification and fermentation process parameters in optimizing bioethanol production

to obtain yields. The negative correction acts as the detrimental effect that affects the production. This AIDMS algorithm uses the correlation coefficient values of every parameter to improve the performance of the AIDMS algorithm. Figure 4 shows the Pearson correlation coefficient values for glucose yield. The cellulose (%) content present in the biomass characteristics is positively correlated to glucose, having a correlation coefficient value of 0.968. This correlation value shows the importance of cellulose in the production of bioethanol. The biomass characteristics such as hemicellulose (%) and lignin (%) are inversely correlated to glucose yield, having coefficient values of -0.698 and -0.883 with ($p > 0.05$). The saccharification process conditions of temperature, acid conc., and time is inversely correlated to glucose yield with a coefficient value of -0.844 , -0.798 and -0.768 with ($p > 0.05$). As time increases, the glucose yield decreases. Except for cellulose (%), all the other biomass characteristics and process conditions have negative correlation values. For saccharification, the cellulose (%) ranges from 0.968 to 0.956, hemicellulose (%) ranges from -0.698 to -0.690 , lignin (%) ranges from -0.877 to -0.883 , acid concentration (%) ranges from -0.794 to -0.798 , temperature ($^{\circ}\text{C}$) ranges from -0.844 to -0.840 , and time (min) ranges from -0.768 to -0.772 .

Fig. 4 Correlation coefficient of glucose analysis using Pearson

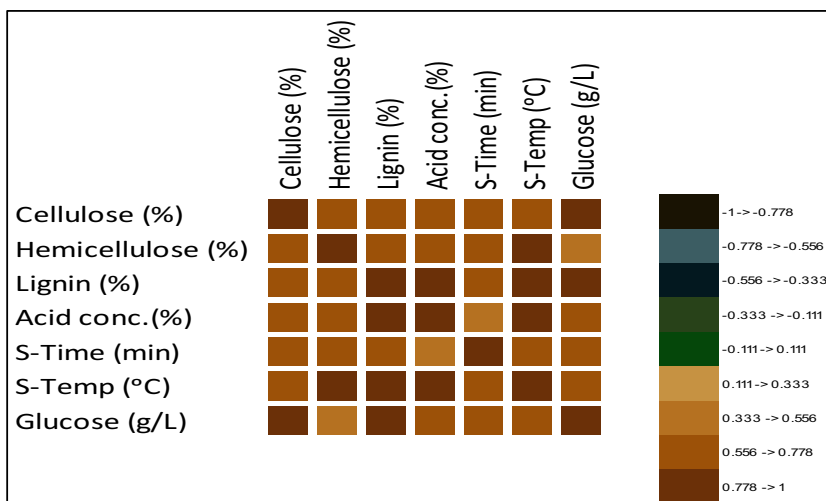


Fig. 5 Correlation coefficient of ethanol yield analysis using Pearson

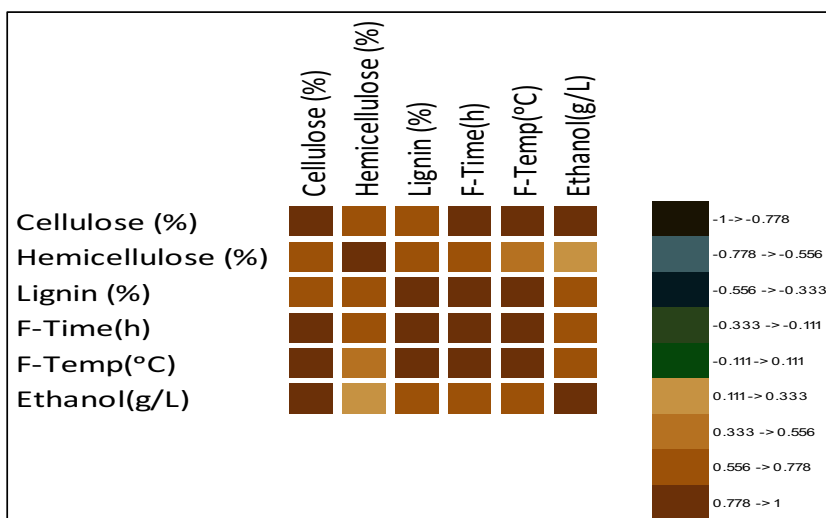


Figure 5 shows the fermentation conditions having a positive correlation coefficient value for cellulose (%) and temperature and time such as 0.904, 0.848, and 0.823 with ($p < 0.05$). The negative correlated coefficient values for hemicellulose (%) and lignin (%) were as follows: -0.551 and -0.753 with ($p < 0.05$). For fermentation, the cellulose (%) ranges from 0.904 to 0.905, hemicellulose (%) ranges from -0.551 to -0.557 , lignin (%) ranges from -0.753 to -0.758 , temperature ($^{\circ}\text{C}$) ranges from 0.845 to 0.848, and time(min) ranges from 0.823 to 0.825.

3.3 Weighted rank order aggregate (WROA)

Considering three biomass characteristics, three saccharification process conditions, and two fermentation process conditions, a total of 11 parameters were given rank order using the weighted rank order aggregate function. Based on the WROA, the first rank was given to cellulose (%) in both the glucose and ethanol yield analyses. Figure 6 shows the

rank order for biomass characteristics and saccharification conditions. Cellulose (%) is considered an important factor with the first rank 232 times in the 250 training dataset. Temperature was ranked second with a weightage of 227 times. The third rank was given to acid concentration (%) with the rank weightage of 224 times. High cellulose content will increase the glucose yield. It provides a good yield over a lower temperature and time. Figure 7 shows the WROA for biomass characteristics and fermentation conditions. Cellulose (%) has taken first place with a weightage of 229 times. Second and third place were given to temperature and time with a weightage of 220 and 219 times. The Euclidean distance method was used in this AIDMS algorithm. It calculates the distance between similar parameter data. A weighted rank was given to the data points having the same distance. If the distance between similar datasets varies, the weighted rank was given to the data according to the nearest neighbor distance. This process was repeated several times, for a total of 250 dataset runs in the WROA algorithm. The

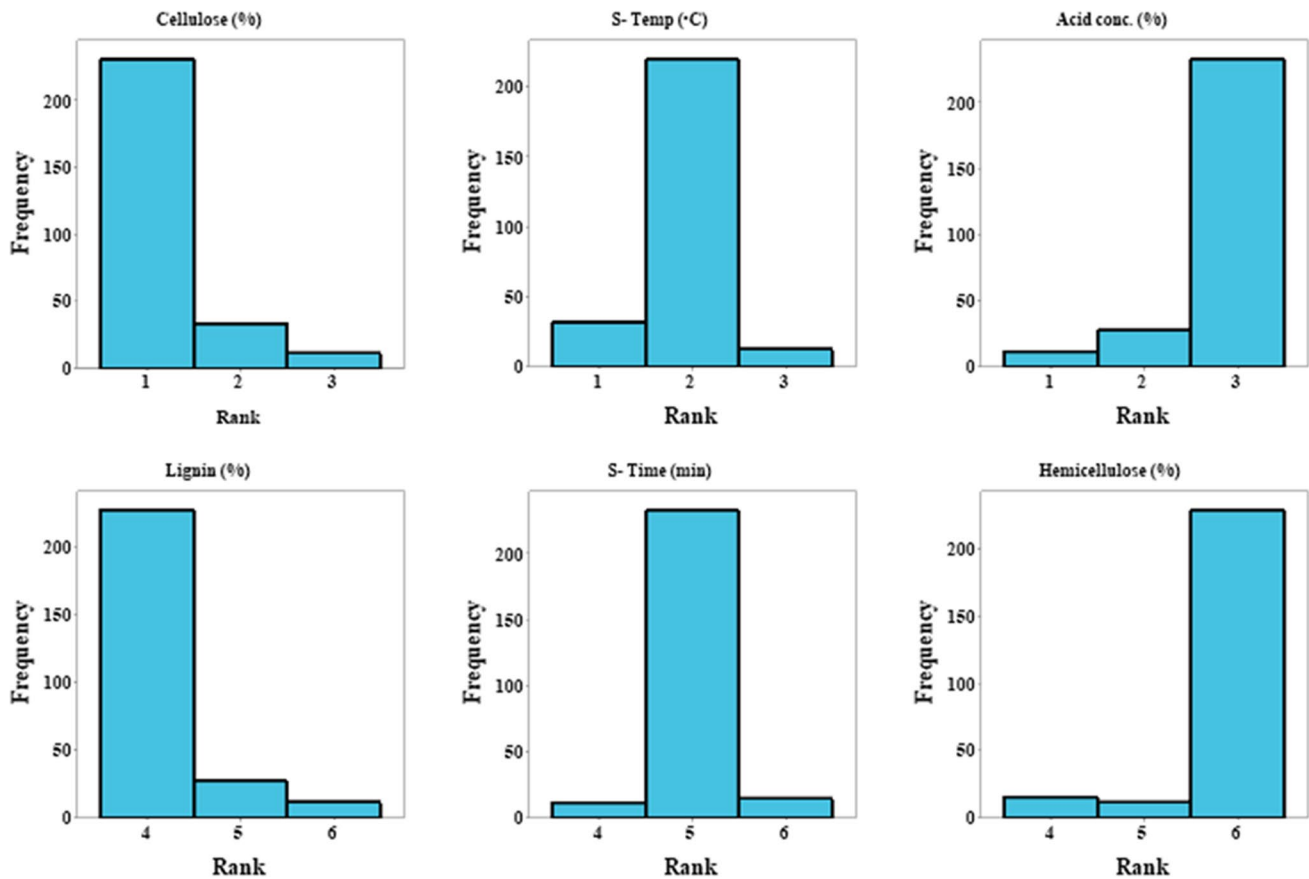


Fig. 6 Frequency histogram plot of WROA for glucose yield

outputs were generated after the number of trials fed to the JB-2 and finally stored in the storage block of the cloud repository.

3.4 Prediction of biomass characteristics and process conditions

The JB-2 block makes the final decision. For optimizing the bioethanol production conditions, 20 datasets were given as the testing data. At the beginning stage of the AIDMS algorithm workflow, the testing dataset containing the biomass characteristics as input, the obtained correlation values are stored in the SB of the cloud repository. Based on their similarity, rank order was given to the JB-2. In JB-2, the optimum process conditions were compared with the stored dataset to calculate the residuals. This process is repeated several times until the best result is attained for the given yield which should be similar to the test data. The predicted data was stored in the SB of the cloud repository as additional training data. So the residual fit is suitable. In this bioethanol production, the glucose and ethanol yield values can be calculated using the AIDMS algorithm.

Figure 8 shows the residue plot for biomass characteristics and process conditions for bioethanol yield. From the figure, it is clearly shown that all the parameters fit well with an accuracy of $> 94\%$. The AIDMS algorithm shows a higher R^2 value and fits well with the experimental data. This algorithm predicts the process conditions by using the biomass name and characteristics as input when the glucose yield is fixed. To evaluate the accuracy of the AIDMS algorithm, by using the coefficient of determination R^2 values were obtained by comparing the 20-test dataset with the predicted AIDMS dataset. The cross-validation step is used to evaluate the performance of the AIDMS model. It ensures the robustness and good generalization of the model chosen for deployment to new data. The performance of the AIDMS algorithm for predicting ethanol production across different feedstock results is effective because it reviews the dataset several times until it attains the desired output. By comparing the AIDMS-predicted dataset with the experimental dataset, 94% accuracy was attained. The AIDMS algorithm can learn from several feedstock compositions. They are strong generalizers to novel feedstock types because they capture fundamental patterns. Because of the variations in process parameter

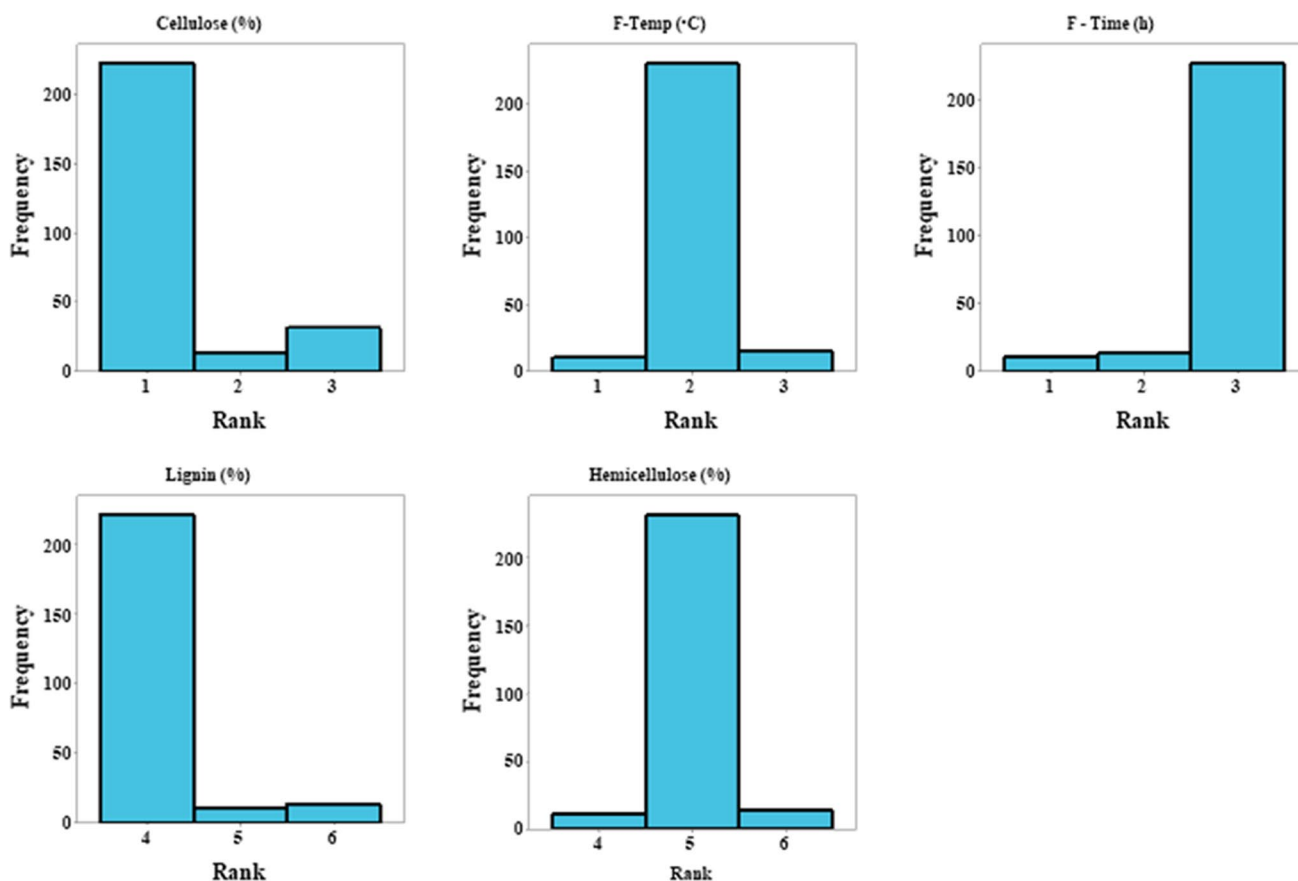


Fig. 7 Frequency histogram plot of WROA for ethanol yield

conditions, this AIDMS algorithm can learn the relationship between the conditions and the bioethanol yield by making them available across various setups, coming to scaling of AIDMS, this algorithm can adapt to industrial scale and lab-scale experiments by improving the accuracy during the validation.

Real-world industrial setting applications include the following:

1. Optimization of dilute acid hydrolysis process parameter to obtain glucose and ethanol yields by using the AIDMS algorithm.
2. Maintenance of the predicted dataset in the cloud repository.
3. Efficiency can be improved by analyzing the process parameter dataset.

Potential impact on bioethanol efficiency includes the following:

1. Maximum glucose and ethanol yields due to optimization of saccharification and fermentation process parameters.

2. Cost can be reduced by preferring this AIDMS algorithm instead of using high-cost equipment.
3. This AIDMS algorithm can adapt to any feedstocks.

The limitations of this study are data collection, data quality, and interpretation of results. The limitations like data collection may affect the validity and scope of the findings by taking time and resources. Similarly, data quality can be improved by supplementing the training dataset by implementing different copies of the minority classes. It can be done many times to improve the accuracy of the results. Similarly, interpretation of results can be reduced by using SVM for classifications and regressions. It improves the prediction of the dataset using the AIDMS algorithm.

3.5 Validation of AIDMS

The biomass characteristics and process conditions predicted by the AIDMS algorithm were validated using experimental results shown in Table 1. Five biomass feedstocks were used for validating the experiments, and their biomass characteristics and process conditions were given as the input in AIDMS. The process conditions were set

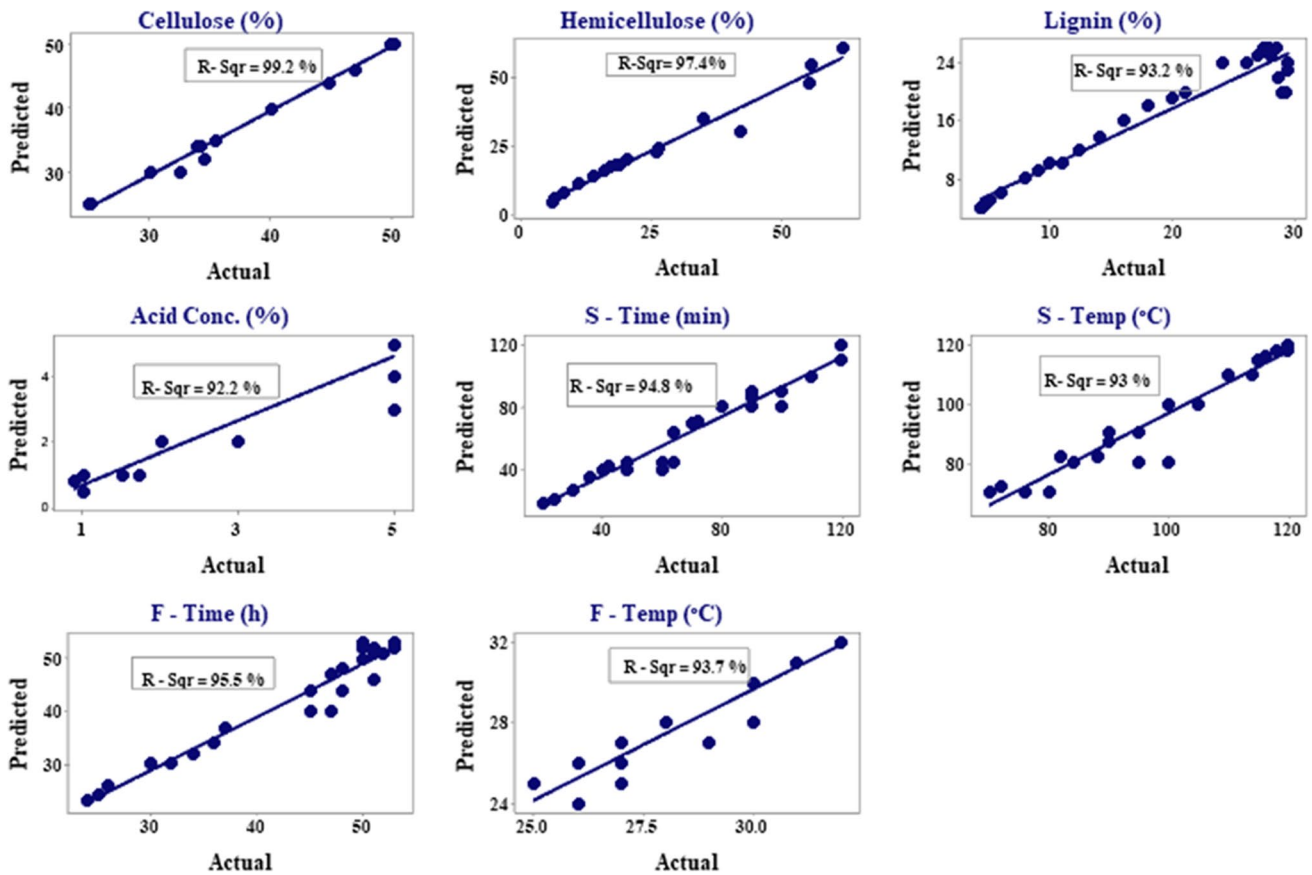


Fig. 8 Comparison of predicted data of acid hydrolysis—saccharification and fermentation conditions from the actual data available in the test dataset

in experiments, and the glucose and ethanol yields were determined. Similarly, the AIDMS also obtains the glucose and ethanol yields for five biomass feedstocks. Accuracy % can be calculated from Eq. 1,

$$\text{Accuracy}(\%) = \left(100 - \frac{\text{Expt.data} - \text{predicted data from AIDMS}}{\text{Experimental data}}\right) \times 100 \quad (1)$$

From Table 1, it is shown that

- I. When the saccharification temperature increased, the glucose yield was reduced.
- II. The AIDMS algorithm has biomass characteristics as a training dataset and the process conditions were predicted for the selected biomass.
- III. If the cellulose is 44.95%, hemicellulose is 55.31%, lignin is 22.71%, and acid concentration is 5%, rice straw waste biomass yielded low glucose and ethanol because of its increase in hemicellulose, lignin, and acid concentration.

IV. Glucose and ethanol yields were low for banana plant waste biomass because of its cellulose content (30.13%) and high hemicellulose (42.01%), lignin (27%), and acid concentration (5%). As the sacchari-

fication time increases, the glucose yield decreases.

3.6 Validation of targeted AIDMS

A targeted AIDMS model predicts the biomass characteristics and process conditions for the glucose and ethanol yields by the user. Five biomass feedstocks such as cotton stalk, wheat straw, olive tree, potato peel waste, rice straw, cotton stalk, and sugarcane bagasse were selected from the various studies for verification of this algorithm. The targeted glucose yield was set to determine the process conditions and ethanol yield. By comparing the targeted glucose yield,

Table 1 Performance of Artificial Intelligence Decision-Making System (AIDMS) based on experimental validation

Input		Process conditions						Output						
Biomass name and organism	Biomass characteristics	Biomass character-istics			Process conditions			Yield						
		C (%)	H (%)	L (%)	Acid conc.	S-Time (min)	S-Temp (°C)	F-Time (h)	F-Temp (°C)	Actual glucose (g/L)	Predicted glucose (g/L)	Actual ethanol (g/L)	Predicted ethanol (g/L)	Accuracy (%)
Cotton stalk and <i>S. cerevisiae</i>		40	13	29	3	30	70	50	30	29.5	29.3	32.8	32	97
Wheat straw and <i>S. cerevisiae</i>		35.6	26.4	13.7	2	70	114	26	25	4.1	3.9	12.4	12	96
Rice straw and <i>S. cerevisiae</i>		44.95	55.31	22.71	5	100	100	32	27	1.6	1.5	9.1	8.6	94
Banana plant waste and <i>S. cerevisiae</i>		30.13	42.01	27	5	110	120	24	26	1.1	1.2	7.2	7	97
Corn cob and <i>S. cerevisiae</i>		50.3	61.6	20	5	120	120	36	28	1.8	2.1	6.1	5.8	95
Olive tree and <i>S. cerevisiae</i>		25.1	18.8	18.5	0.89	50	84	45	31	20.6	19.2	35.1	33	94
Potato peel waste and <i>S. cerevisiae</i>		34.3	6	4.3	1	100	120	36	28	1.3	1.3	21.7	21	97
<i>Curcuma longa</i> and <i>S. cerevisiae</i>		51	11	12	2	20	70	55	32	31.4	31.4	54.7	54.3	99
Giant reed and <i>S. cerevisiae</i>		34.6	20.4	20.4	2	30	70	52	32	30.8	30.8	32.5	31	95
Potato peel waste and <i>S. cerevisiae</i>		34.3	6	4.3	5	110	120	48	29	1.8	1.8	28.6	27	94

the results obtained are > 94%. Out of 250 acid hydrolysis biomass datasets, the training dataset contains the collected dataset as well as the AIDMS model-predicted dataset so the accuracy of the training dataset is higher than the test dataset accuracy [54]. The accuracy of the training dataset obtained using this AIDMS algorithm is 96% which is more compared to the accuracy of the test dataset 94%. This proves that the Artificial Intelligence Decision-Making System performs well in this optimization of the dilute acid hydrolysis process.

4 Conclusion

This study uses machine learning to develop an algorithm called the Artificial Intelligence Decision-Making System to optimize the production of bioethanol from different biomass feedstocks. A hundred datasets act as training data and 20 as test data. The Pearson correlation coefficient matrix shows the correlation between glucose and ethanol yields for biomass characteristics and process conditions. A weighted rank order aggregate was given to find the important factor in determining the yield. The accuracy % was determined using actual values and values predicted using the AIDMS algorithm and found to be more than 94%. For feedstock such as cotton stock, banana plant waste, 97% of accuracy, and for wheat straw 96% was attained through this algorithm. This artificial intelligence-based system plays an important role in optimizing bioethanol production. In the future, using this AIDMS algorithm, a master data management system was developed for bioethanol production for implementation in the industry. It can also be implemented to optimize Hydro Thermal Liquefaction and Hydro Thermal gasification.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13399-024-05714-y>.

Author contribution Nithianantharaj Vinitha: design of experiments, experimentation, analysis of data; Jaikumar Vasudevan, K.P. Gopinath: writing of the manuscript and investigation; J. Arun: curated data from the study; S Naveen, S Madhu: validation of data, review of the manuscript.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Nil

Consent to participate Nil

Consent for publication Nil

Competing interests The authors declare no competing interests.

References

- Duarte GT, De Alencar Nääs I, Da Silva Lima ND (2021) Estimating the urban environmental impact of gasoline-ethanol blended fuels in a passenger vehicle engine. *Environmental Science and Pollution Research*, pp 1–12
- Bulus GC, Koc S (2021) The effects of FDI and government expenditures on environmental pollution in Korea: the pollution haven hypothesis revisited. *Environmental Science and Pollution Research* 28(28):38238–38253
- Mele M, Magazzino C (2021) Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence. *Environ Sci Pollut Res* 28:2669–2677
- Demiray E, Karatay SE, Dönmez G (2019) Improvement of bioethanol production from pomegranate peels via acidic pretreatment and enzymatic hydrolysis. *Environmental Science and Pollution Research* 26:29366–29378
- Casabar JT, Unpaprom Y, Ramaraj R (2019) Fermentation of pineapple fruit peel wastes for bioethanol production. *Biomass Conversion and Biorefinery* 9(4):761–765
- Osman ME, Abo-Shady AM, Elshobary ME, Abd El-Ghafar MO, Abomohra AE-F (2020) Screening of seaweeds for sustainable biofuel recovery through sequential biodiesel and bioethanol production. *Environmental Science and Pollution Research* 27:32481–32493
- Binod P, Sindhu R, Singhanian RR, Vikram S, Devi L, Nagalakshmi S, Kurien N, Sukumaran RK, Pandey A (2010) Bioethanol production from rice straw: an overview. *Bioresour Technol* 101(13):4767–4774
- Duque A, Álvarez C, Doménech P, Manzanares P, Moreno AD (2021) Advanced bioethanol production: from novel raw materials to integrated biorefineries. *Processes* 9(2):206
- Long F, Liu H (2023) An integration of machine learning models and life cycle assessment for lignocellulosic bioethanol platforms. *Energy Conv Manag* 292:117379
- Xu Z, Huang F (2014) Pretreatment methods for bioethanol production. *Applied biochemistry and biotechnology* 174(1):43–62
- Sahu S, Pramanik K (2018) Evaluation and optimization of organic acid pretreatment of cotton gin waste for enzymatic hydrolysis and bioethanol production. *Applied biochemistry and biotechnology* 186(4):1047–1060
- Demiray E, Kut A, Karatay SE, Dönmez G (2021) Usage of soluble soy protein on enzymatically hydrolysis of apple pomace for cost-efficient bioethanol production. *Fuel* 289:119785
- Kumar AK, Parikh BS, Pravakar M (2016) Natural deep eutectic solvent mediated pretreatment of rice straw: bioanalytical characterization of lignin extract and enzymatic hydrolysis of pretreated biomass residue. *Environmental Science and Pollution Research* 23:9265–9275
- Tulashie, SK, Akpari, EEA, Appiah, G, Adongo, A & Andoh, EK 2021, 'Acid hydrolysis of sawdust waste into bioethanol', *Biomass Conversion and Biorefinery*, pp. 1–14.
- Slathia PS, Raina N, Kiran A, Kour R, Bhagat D, Sharma P (2020) Dilute acid pretreatment of pine needles of *Pinus roxburghii* by response surface methodology for bioethanol production by separate hydrolysis and fermentation. *Biomass Conversion and Biorefinery* 10(1):95–106
- De Farias Silva CE, Bertuccio A (2017) Dilute acid hydrolysis of microalgal biomass for bioethanol production: an accurate kinetic model of biomass solubilization, sugars hydrolysis and nitrogen/ash balance. *Reaction Kinetics, Mechanisms and Catalysis* 122(2):1095–1114
- Dai Z, Chen Z, Selmi A, Jermittiparsert K, Denić NM, Nešić Z (2021) Machine learning prediction of higher heating value of biomass. *Biomass Convers Biorefin*, 1–9

18. Pereira LMS, Milan TM, Tapia-Blácido DR (2021) Using response surface methodology (RSM) to optimize 2G bioethanol production: a review. *Biomass and Bioenergy* 151:106166
19. Seo MW, Lee SH, Nam H, Lee D, Tokmurzin D, Wang S, Park Y-K (2022) Recent advances of thermochemical conversion processes for biorefinery. *Bioresource technology* 343:126109
20. Milić M, Petković B, Selmi A, Petković D, Jermsittiparsert K, Radivojević A, Milovancević M, Khan A, Vidosavljević ST, Đenić N (2021) Computational evaluation of microalgae biomass conversion to biodiesel. *Biomass Convers Biorefin*, 1–8
21. Owusu WA, Marfo SA (2023) Artificial intelligence application in bioethanol production. *Int J Energy Res* 2023
22. Momenitabar M, Ebrahimi ZD, Abdollahi A, Helmi W, Bengtson K, Ghasemi P (2023) An integrated machine learning and quantitative optimization method for designing sustainable bioethanol supply chain networks. *Decis Analytics J* 7:100236
23. Zhang Y, Li R&, Zhang J (2021) Optimization scheme of wind energy prediction based on artificial intelligence. *Environmental Science and Pollution Research* 28:39966–39981
24. Smuga-Kogut M, Kogut T, Markiewicz R, & Słowik A (2021) Use of machine learning methods for predicting amount of bioethanol obtained from lignocellulosic biomass with the use of ionic liquids for pretreatment. *Energies* 14(1):243
25. Huntington T, Cui X, Mishra U, Scown CD (2020) Machine learning to predict biomass sorghum yields under future climate scenarios. *Biofuels, Bioproducts and Biorefining* 14(3):566–577
26. Kylili K, Kyriakides I, Artusi A, Hadjistassou C (2019) Identifying floating plastic marine debris using a deep learning approach. *Environmental Science and Pollution Research* 26:17091–17099
27. Palanisamy A, Soundarrajan N, Ramasamy G (2021) Analysis on production of bioethanol for hydrogen generation. *Environ Sci Pollut Res* 28(45):63690–63705
28. Petković B, Petković D, Kuzman B (2022) Adaptive neuro fuzzy predictive models of agricultural biomass standard entropy and chemical exergy based on principal component analysis. *Biomass Convers Biorefin* 12(7):2835–2845
29. Gundupalli MP, Cheng Y-S, Chueter S, Bhattacharyya D, & Sriariyanun M (2021) Effect of dewaxing on saccharification and ethanol production from different lignocellulosic biomass. *Bioresource technology* 339:125596
30. Sundarrajan P, Gopinath KP, Arun J, Gracepavithra K, Pavendan K, Adithyayoseph A (2020) An insight into carbon balance of product streams from hydrothermal liquefaction of *Scenedesmus abundans* biomass. *Renewable Energy* 151:79–87
31. Świątek K, Gaag S, Klier A, Kruse A, Sauer J, Steinbach D (2020) Acid hydrolysis of lignocellulosic biomass: Sugars and furfurals formation. *Catalysts* 10(4):437
32. Sriariyanun M, Amnuaycheewa P, Rodiahwati W, Sanvarinda PP, Cheenkachorn K, & Tawai A (2017) Effect of organic acid pretreatment on Napier grass (*Pennisetum purpureum*) straw biomass conversion. *10:2Applied Science and Engineering Progress*
33. Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal Chem* 31(3):426–428
34. Caputi A, Ueda M, Brown T (1968) Spectrophotometric determination of ethanol in wine. *Am J Enol Viticult* 19(3):160–165
35. Bansal M, Goyal A, Choudhary A (2022) A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decis Analytics J* 3:100071
36. Victoria AH, Maragatham G (2021) Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems* 12(1):217–223
37. Cho H, Kim Y, Lee E, Choi D, Lee Y, Rhee W (2020) Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks. *IEEE access* 8:52588–52608
38. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H (2019) Hyperparameter optimization for machine learning models based on bayesian optimization. *J Electron Sci Technol* 17(1):26–40
39. Prakash KB, Kanagachidambaresan G (2021) Programming with TensorFlow
40. Hao J, Ho TK (2019) Machine learning made easy: a review of scikit-learn package in python programming language. *J Educational Behav Stat* 44(3):348–361
41. Zubaedah R, Xaverius F, Jayawardana H, Hidayat SH (2020) Comparing euclidean distance and nearest neighbor algorithm in an expert system for diagnosis of diabetes mellitus. *Enfermeria clinica* 30:374–377
42. Lenihan P, Orozco A, O’neill E, Ahmad M, Rooney D, Walker G (2010) Dilute acid hydrolysis of lignocellulosic biomass. *Chem Eng J* 156(2):395–403
43. Zhou Z, Liu D&, Zhao X (2021) Conversion of lignocellulose to biofuels and chemicals via sugar platform: an updated review on chemistry and mechanisms of acid hydrolysis of lignocellulose. *Renew Sustain Energy Rev* 146:111169
44. Horn SJ, Vaaje-Kolstad G, Westereng B, Eijsink V (2012) Novel enzymes for the degradation of cellulose. *Biotechnol Biofuels* 5(1):1–13
45. Das N, Jena PK, Padhi D, Kumar Mohanty M, Sahoo G (2021) A comprehensive review of characterization, pretreatment and its applications on different lignocellulosic biomass for bioethanol production. *Biomass Convers Biorefin*, 1–25
46. Dussán KJ, Silva D, Moraes E, Arruda PV, Felipe M (2014) ‘Dilute-acid hydrolysis of cellulose to glucose from sugarcane bagasse’, *Chemical Engineering Transaction*, 38
47. Fan L-T, Gharpuray MM, Lee Y-H (1987) Acid hydrolysis of cellulose. *Cellulose hydrolysis*. Springer, pp 121–148
48. Huang L-Z, Ma M-G, Ji X-X, Choi S-E, Si C (2021) Recent developments and applications of hemicellulose from wheat straw: a review. *Front Bioeng Biotechnol* 9:690773
49. Pessoa A Jr, Mancilha I, & Sato S (1997) ‘Acid hydrolysis of hemicellulose from sugarcane bagasse’. *Braz J Chem Eng* 14
50. El Harchi M, Kachkach FF, El Mtili N (2018) Optimization of thermal acid hydrolysis for bioethanol production from *Ulva rigida* with yeast *Pachysolen Tannophilus*. *South Afr J Bot* 115:161–169
51. Keshav PK, Shaik N, Koti S, Linga VR (2016) Bioconversion of alkali delignified cotton stalk using two-stage dilute acid hydrolysis and fermentation of detoxified hydrolysate into ethanol. *Industrial Crops and Products* 91:323–331
52. Nguyen CM, Nguyen TN, Choi GJ, Choi YH, Jang KS, Park Y-J, Kim J-C (2014) Acid hydrolysis of *Curcuma longa* residue for ethanol and lactic acid fermentation. *Bioresour Technol* 151:227–235
53. Qi B, Vu A, Wickramasinghe SR, Qian X (2018) Glucose production from lignocellulosic biomass using a membrane-based polymeric solid acid catalyst. *Biomass and Bioenergy* 117:137–145
54. Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. *Catena* 145:164–179

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.