Biochemical Engineering Journal 181 (2022) 107896

Contents lists available at ScienceDirect

Biochemical Engineering Journal

journal homepage: www.elsevier.com/locate/bej

Ensemble models of feedstock blend ratios to minimize supply chain risk in bio-based manufacturing

Chyi-Shin Chen [1], Akash Narani, Aigerim Daniyar [2], Joshua McCauley [3], Sarah Brown [4], Todd Pray [5], Deepti Tanjore *

Advanced Biofuels and Bioproducts Process Development Unit (AB-PDU), Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States

ARTICLE INFO

Keywords:
Feature selection
Feedstock blends
Lignocellulosic biomass
Machine learning
Predictive model

ABSTRACT

Feedstock blending as a strategy to mitigate risks in the supply of lignocellulosic biomass to commercial scale biorefineries across various geographical areas in the United States. Machine learning predictive models estimate sugar yields from feedstock blends expected to be available in Florida and Kansas. Performance of each model was assessed based on feature selection along with two-stack ensembles applied in both linear and nonlinear algorithms. Linear-weighted ensemble and nonlinear stochastic gradient boosting model ensembled with four base learners exhibited similar predictions as previously developed linear regression models when predicting glucose yields in the higher range. The ensemble models achieved a 10–50 % improvement in the root mean squared error with feature selection compared to models with full features from validation. Machine learning has the potential to predict sugar yields at high confidence for a given feedstock blend ratio and pretreatment conditions.

1. Introduction

The conversion of lignocellulosic feedstocks to long-chained alkanes and other energy-dense molecules is a promising path to minimizing dependence on crude oil for jet and other transportation fuels [1,2]. According to the 2016 Billion-Ton report, the United States (US) can generate 702 million dry tons of biomass every year and has the potential of generating more than 1 billion dry tons of biomass by 2030 at the farm gate price of $60/dry ton [3]. However, it is important to note that the biomass available in the US is varying, ranging from agricultural and forest residues to energy crops. To operate a commercial biorefinery year round, over 2205 dry US tons of biomass feedstock needs to be processed per day, for 330 days [4]. However, most geographical areas in the US do not offer a single feedstock at such magnitudes. In order to overcome the feedstock variability and to expand bio-based manufacturing, it is critical to develop strategies that can be applied to variegated biomass types available in the vast geographical expanse of the US.

Most researchers studying lignocellulosic feedstock conversion, however, do not study feedstock blends and instead focus on a single biomass feedstock, often corn stover [5–8]. The recent collaborative efforts between Lawrence Berkeley and Idaho National Laboratories (INL) have identified geographical locations where feedstock blending can commence bio-based processing at commercial scale. We studied the availability of feedstocks in western Florida and developed a predictive model that identified optimal blend ratios of corn stover (CS) - a high-quality easily convertible feedstock - with local low-quality feedstocks: energy cane (EC) and switchgrass (SG) [9]. CS would be transported from Georgia to Florida, as it can ensure the complete utilization of local feedstocks. The optimized blends along with associated pretreatment conditions to maximize sugar (glucose) yields and thereby fuel titers were presented as a strategy to mitigate supply risks and feedstock variability concerns at commercial-scale biorefineries in geographical locations that do not have abundant access to a single

* Corresponding author.
E-mail address: DTanjore@lbl.gov (D. Tanjore).
[1] Current Affiliation: Yamaguchi University, Yamaguchi, Japan.
[2] Current Affiliation: School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR.
[3] Current Affiliation: Joint BioEnergy Institute and Agile BioFoundry Consortium, Lawrence Berkeley National Laboratory, Berkeley, CA.
[4] Current Affiliation: University of Melbourne, Melbourne, Australia.
[5] Current Affiliation: Strategic Partnership Office, Lawrence Berkeley National Laboratory, Berkeley, CA.

https://doi.org/10.1016/j.bej.2020.107896
Received 20 July 2020; Received in revised form 13 December 2020; Accepted 15 December 2020
Available online 24 December 2020
1369-703X/

feedstock [10].

Glucose yields from lignocellulosic biomass pretreatments have a significant impact on the economic viability in bio-based process chains. Glucose yields depend on feedstock compositions and multiple pretreatment process parameters, and testing all combinations of these parameters with each feedstock blend is both resource and time constraining [11–14]. Accurately predicting glucose yields based upon changing feedstock blends and process variables thus can be very useful, especially in a biorefinery setting, where 2205 dry tons of variable feedstock is received on a daily basis. The predictive model can help optimize process conditions to convert the varying feedstock in real-time. Linear regression models using SAS JMP® (Cary, NC) from our previous study [9] could assess the impact of feedstock composition and each of the parameters and predict glucose yields. While the impact of parameters was thoroughly established, our predictions exhibited a root mean squared error (RMSE) of 8.01.

To improve prediction performance, in this study, we developed ensemble machine learning (ML) models and applied them on this first dataset obtained from feedstocks available in Florida (dataset 1). Although ML has been widely adopted in many fields of research, including other bio-related fields [15–17], there have been very few applications of ML in deconstruction studies involving either a single or multiple feedstocks [18–20]. In this study, we tested the application of ML models such as single regression and 2-fold ensemble models integrated with four base learners in linear-weighted regression and nonlinear stochastic gradient boosting models (gbm). Fig. 1 illustrates the architecture of the layers of the ensemble models, which include base models (layer 0) and ensemble model (layer 1). Starting from data processing steps, four base models were selected from eighteen model candidates. Fine tuning feature selection technique using decision trees was applied to the data to improve model performance, which was measured by RMSE using glucose yield from experimental studies. To ascertain the applicability of the ML methods, we further tested the methods on a second dataset from blends of CS and SG with wheat straw (WS), feedstock combinations expected in the state of Kansas (dataset 2). By developing layers of ensemble models as shown in Fig. 1, the predictive power of the ML model for glucose yield was confirmed and the role played by each feedstock in blends was also revealed by examining features using an extreme gradient boosting model.

2. Materials and methods

2.1. Biomass feedstocks, pretreatment catalysts, and other experimental methods

INL (Idaho Falls, ID) supplied feedstocks to generate dataset 1 (CS, SG, EC) and dataset 2 (CS, SG, WS). INL's least cost formulation (LCF) model was used to determine feedstocks for testing. We chose two US states outside of the corn belt, Florida and Kansas, which enjoy an abundant supply of locally grown feedstock such as EC in Florida and WS in Kansas. These feedstocks were tested for integration into the local supply chain by blending with higher quality feedstocks, such as CS and SG. More than 500,000 dry tons of EC along with 100,000 –300,000 dry tons of SG will be available annually in Lee County, Florida by 2030. CS needed to obtain optimal feedstock blends will have to be shipped from other parts of the US. The cost of transportation per different modes were also listed in Narani et al., 2019. As per LCF, the estimated prices for EC, SG, and CS on a dry basis were $70, $50, and$60/ton respectively (Narani et al. 2017). All feedstocks in Sheridan county, KS, listed in the dataset 2 (WS, SG, CS) were abundantly available in the local area and the expected price of the feedstock blends ranged between $81.38 to $84.14/ ton. About 50, 30, and 10 million dry tons of SG, CS, and WS will be available annually in Sheridan country, KS.

Traditional pretreatment catalysts (dilute acid, dilute alkali, and ionic liquid (IL)) in the form of 1% (w/w) sulfuric acid in water, 1% (w/w) sodium hydroxide in water and 99 % purity 1-ethyl-3-methyl-imadozolium acetate were used. Each of the pretreatment catalysts was defined as a categorical variable, such that the experimental design can choose either dilute acid dilute alkali or IL but not the combination of two or three catalysts. Pretreatment reaction temperature and time were applied as scaled variables with unique operating ranges for each pretreatment catalyst. The operating ranges were selected after a comprehensive review of literature for each of the three pretreatment catalysts [9]. These inputs were fed into the SAS JMP® to generate the experimental design listed in Narani et al. [9]. Listed below are the variables used in the experiment.

- Feedstock compositions (scaled): EC (0–100 %), SG (0–100 %), and CS (0–100 %)
- Pretreatment catalyst (categorical): dilute acid, dilute alkali, or IL

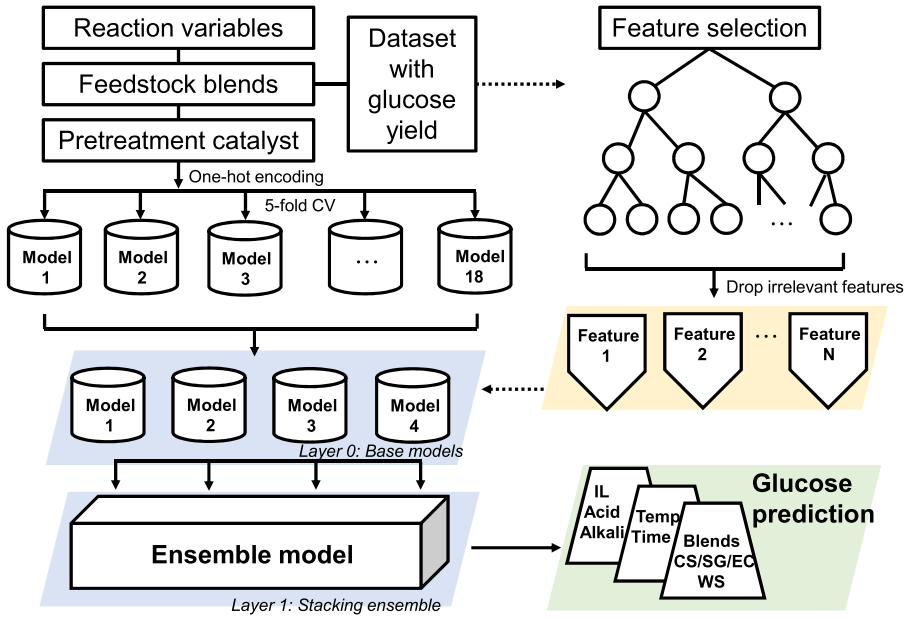

Fig. 1. Model architecture including model selection, feature filtration, and model stacking. 5-fold cross validation (CV) was set for each model during training. The final ensemble was constructed by two layers (layer 0 and layer 1).

- Pretreatment reaction temperatures (scaled): dilute acid (1–100 %) 140–180 °C, dilute alkali (1–100 %) 55–120 °C, and IL (1–100 %) 120–160 °C.
- Pretreatment reaction times (scaled): dilute acid (1–100 %) 5–60 min, dilute alkali (1–100 %) 1–24 h, IL (1–100 %) 1–3 h.

The process variables in datasets 1 and 2 were defined as parameters: (i) pretreatment catalysts (acid, alkali, and ionic liquid), (ii) feedstock blends (Florida Blend: CS, SG, and EC and Kansas Blend: CS, SG, and WS), and (iii) reaction variables (temperature and reaction time). Solid-liquid separation and washing on biomass was conducted after IL pretreatment, prior to saccharification xylan/xylose wass lost during the washing of IL pretreated biomass. While solid-liquid separation after acid and alkali pretreatments was not conducted, to be consistent with our approach and not use xylose yields, only overall glucose yields after saccharification were used for the predictive model development. More details of each experiment can be found in Narani et al. [9] and Table S1.

2.2. Linear regression model and the associated predictions

Our previous manuscript presented a linear mixed effect model for glucose yields by performing data analysis with Restricted Maximum Likelihood, or REML, a method that was available through SAS JMP® [9]. This Linear Regression Model (LRM) and its predictions from dataset 1 were presented in detail in Narani et al. [9]. A total of 74 experimental data points were available in dataset 1 with an additional 13 experimental data points generated to validate LRM performance. The experimental design of dataset 2 was also generated in SAS JMP®, and 74 conditions representing the various combinations of parameters described in Section 2.1 were tested. An additional 16 points were generated to validate the LRM on dataset 2.

For both dataset 1 and 2, the ratios of the three feedstocks had a significant impact on glucose yields (p-value < 0.0001). A larger concentration of recalcitrant feedstock led to lower sugar yields and vice-versa. Reaction temperature and times, within the chosen ranges, had no significant impact on glucose yield for both datasets (p-values of 0.31 and 0.25 for dataset 1 and 0.65 and 0.64 for dataset 2). We believe that our pre-determined experimental range of treatment conditions, which lied within optimal ranges for each of the pretreatment catalysts, was too narrow to find a significant impact. IL catalyst led to very high conversion yields in most tests, leading to statistically significant impact (p-value < 0.0001) on sugar yields in most treatment conditions. The alkali and acid pretreatments had lower but significant impact on glucose yields in dataset 1 (p-values at 0.0005 and 0.042, respectively). Interestingly, while the alkali pretreatment had a significant impact on glucose yields from dataset 2 (p-value < 0.0001), the acid treatment had no significant impact on the same (p-value at 0.092).

2.3. Machine learning methods

2.3.1. Data input

Same data features from both datasets, including validation tests, were used as data input for the machine learning base models. All models were constructed with a 5-fold cross validation. Two-thirds of the data was used for training and the rest were used for testing to examine prediction performance. The first 10 process and output parameters for both the datasets are reported in Table 1.

All ML modeling and analyses were performed using R with related packages (caret, caretEnsemble, and elasticnet). Eighteen ML models were built to screen for the model with the lowest Pearson correlation (Fig. S1 and S2). Models with low RMSE were chosen as base models for a linear weighted ensemble or a gbm model as the final model in the two-fold ensemble method.

2.3.2. Linear weighted ensemble

Linear weighted ensembles find a linear combination of base models and instead of taking an average of their predictions, they apply different weights to each of them, per their relative contribution to the model. The weights are calculated by developing a generalized linear model, which provides coefficients to each model. When the number of models is equal to i, the weight for each individual can be described as

$$W_i = \frac{|C_i|}{\sum_i |C_i|} \quad (1)$$

Table 1
A glimpse of two experimental datasets for modeling process variables on glucose yield.

Process Parameters						Output
Dataset 1						
Pretreatment	Feedstock Ratios			Temperature	Time	Glucose Yield
	Energy Cane	Switch Grass	Corn Stover	°C	min	% Theoretical
Ionic Liquid	0.00	1.00	0.00	120	106.8	49.20
Dilute Acid	0.30	0.40	0.30	140	60.0	54.75
Dilute Alkali	1.00	0.00	0.00	55	1440.0	56.54
Dilute Alkali	0.50	0.50	0.00	55	589.0	33.96
Dilute Alkali	0.00	1.00	0.00	55	1440.0	57.82
Ionic Liquid	1.00	0.00	0.00	120	180.0	52.33
Dilute Acid	0.40	0.60	0.00	180	5.0	27.21
Dilute Acid	0.00	0.00	1.00	180	38.0	57.33
Dilute Alkali	0.00	1.00	0.00	120	60.0	56.78
Dilute Alkali	0.00	0.00	1.00	120	60	74.60
Dataset 2						
Pretreatment	Feedstock Ratios			Temperature	Time	Glucose Yield
	WheatStraw	Switch Grass	Corn Stover	°C	min	% Theoretical
Dilute Acid	0.00	0.07	0.93	140	26.45	74.63
Ionic Liquid	0.34	0.33	0.34	120	180	86.93
Ionic Liquid	0.00	1.00	0.00	120	180	90.00
Dilute Acid	0.50	0.50	0.00	140	26.45	64.92
Dilute Acid	0.50	0.00	0.50	140	60	66.92
Dilute Alkali	0.00	0.00	1.00	55	1440	87.45
Dilute Acid	1.00	0.00	0.00	140	26.45	63.27
Dilute Acid	0.00	1.00	0.00	140	60	54.93
Dilute Alkali	0.00	1.00	0.00	55	598.20	42.14
Dilute Alkali	1.00	0.00	0.00	55	598.20	24.12

where W stands for weight and C represents a coefficient. By applying the weights to each model, the blended prediction of a linear weighted ensemble is in the form of

$$P_{\text{ensemble}}(x) = \sum_i W_i P_i(x) \tag{2}$$

where P is the prediction value from a model and x represents each data point in the data used for the model. Although nonlinear algorithms are usually good at blending models and incorporating features, linear regression is the most widely used algorithm. It requires less tuning and the result is easy to understand for most users. For these reasons, a linear weighted ensemble method was tested in this study.

2.3.3. Stochastic gradient boosting

Stochastic gradient boosting was introduced by Friedman [21,22] and became known for being a non-linear algorithm that improves a single weak learner to a stronger one by sequential training. The minimum mean squared error (MSE) of each model is improved by introducing randomness, which includes taking subsets of the full data set and randomly applying the subsets as training data. Iterations of the model are updated with new data subsets until MSE cannot be reduced any further. By incorporating bagging to the function estimation process, the randomness could further improve the model performance [23]. The concept can be expressed as

$$F_m(x_i) = F_{m-1}(x_i) + \beta_m h(x_i; a_m) \tag{3}$$

where m is the number of stages to train the models, F is the function in the model, x represents each data point in the data in the order of i, β is the expansion coefficient, and h stands for a regression tree of x with parameters a.

The model development is initiated at the average of the output value (y), which is glucose yield in this study:

$$F_0(x_i) = \text{argmin} \sum_i L(y_i, \gamma) \tag{4}$$

where γ is defined as pseudo residuals. For models from 1 to m,

$$\gamma_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{5}$$

where $F(x)$ is estimated by minimizing $L(y, \gamma)$, where $h(x;a)$ is used to solve L function by fitting to least squared errors. The coefficient β can be determined by a given $h(x;a)$ as

$$\beta_m = \text{argmin} \sum_i L(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)) \tag{6}$$

To prevent overfitting, the shrinkage parameter is applied as 0.1 in each iteration to control the learning rate. By computing residuals from the model in each stage, a regression tree is fit to the residual and added to the next iteration, which is accomplished by gbm algorithms in R. The advantages of stochastic gradient boosting include low sensitivity to outliers, good performance for unbalanced data, and an increase in robustness, accuracy, and execution speed by introducing randomness, which makes it ideal for this study.

2.3.4. Model performance

The final ensemble has a first layer of four base learner models with the low inter-model correlations and RMSE, and a second layer model in linear weighted or gbm model. Base models were chosen based on linear correlation among each other by Pearson correlation coefficient (r) as given by Eq. (7) and RMSE. The models were compared based on the predictive performance using test and validation data from datasets 1 and 2. The performance was evaluated by RMSE in glucose yield, and coefficient of determination (R^2) and mean absolute error (MAE) were examined as well. Each value was calculated according to Eqs. (8),(9), and (10) respectively.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_i^n (y_{\text{exp},i} - y_{\text{pred},i})^2}{n}} \tag{8}$$

$$R^2 = 1 - \frac{\sum_i^n (y_{\text{exp},i} - y_{\text{pred},i})^2}{\sum_i^n (y_{\text{exp},i} - \overline{y_{\text{exp}}})^2} \tag{9}$$

$$\text{MAE} = \frac{\sum_i^n |y_{\text{exp},i} - y_{\text{pred},i}|}{n} \tag{10}$$

where n is the number of points, x_i *and* y_i are individual value from models being compared, $\bar{x}$ and $\bar{y}$ are the mean of predictive glucose yield from models, $y_{\text{exp},i}$ is the actual glucose yield, $y_{\text{pred},i}$ is the predicted glucose yield, and $\overline{y_{\text{exp}}}$ is the mean of actual glucose yield.

The significance of each process variable was measured by an extreme gradient boosting model, and features with significance over 20 (importance level) were used in feature-selection models. The performance in predicting glucose yield was compared with models using all features.

3. Results and discussion

3.1. Drawbacks of predictions from linear regression

Our LRM was able to statistically identify the parameters that had the most impact on glucose yields. However, during validation, we observed that the error in our predictions was high especially in dataset 2 (RMSE for dataset 1 was 8.01 with R^2 of 0.85, and for dataset 2 was 14.65 with R^2 of 0.68) as we used only a single base model that led to overfitting of the training data and contributed to low prediction accuracy. In the LRM model, we only studied the main effects from feedstock blend ratios, pretreatment catalysts, reaction temperature and time and not the interactions among these test parameters. Since this was the first known attempt of building a predictive model on lignocellulosic feedstock blends, simplicity and reliable understanding were our reasons to pursue only main effects. But, in order to better explain our data, we realized that we need ML models with more control on the training levels / algorithms that will offer more flexibility and improve prediction performance. LRM was inadequate, mostly because of limited data and complex process input variables. Two types of data: discrete (e.g. type of pretreatment catalyst) and continuous (e.g. feedstock blend ratio and pretreatment time) were essentially difficult to discern in the linear methods and unable to exlain the relationship between dependents and independents. Nonlinear models, such as multiple degree polynomials or decision tree algorithms can reflect the most important characteristic. For example, a regression tree model accepts the pretreatment time and temperature variables as continuous values with in-built feature selection and returns the predictions that are ascertained after judging through multiple leaves as shown in Fig. S3. As a result, it was imperative that we explore non-linear models such as decision trees, which are better able to discern among the distinct pretreatment catalysts and better predict the glucose yield from the dependent and independent features of the dataset.

3.2. Model selection

To choose the best model for predicting glucose yield, eighteen models including linear and nonlinear types of regression analysis were constructed on the same training dataset. RMSE was used as the goodness of fit for each model and inter-model correlation was examined. Some of the models had Pearson correlation coefficients ranging

between 0.3 and 0.6, while others had correlation coefficients more than 0.9. RMSEs were similar among most of the models, as shown in Fig. 2.

Models with correlation coefficients below 0.9 were selected from Fig. S1 and S2, and the models with low inter-model correlations or RMSE were chosen as base learners to maximize accuracy and minimize prediction correlation for the next layer of ensemble. For dataset 1, Bayesian ridge regression (bridge), an extension of Quinlan's M5 model tree (cubist), projection pursuit regression (ppr), and regression trees (rpart) were selected. As for dataset 2, Breiman's random forest algorithm (rf) and extreme gradient boosting (xgbTree) were chosen along with cubist and ppr.

The Pearson coefficients combined with the scatter plot of the prediction points from selected models are illustrated by the scatterplot matrix, which is known as SPLOM, in Fig. 3. First invented by John Hartigan [24], SPLOM is composed of multiple scatter plots of computed / predicted data (e.g. glucose yield in this study) between models. The symmetric matrix of paired scatter plots allows an easy way to conceptualize the potential correlations between models or variables. In the bottom part of Fig. 3, scatterplots of predicted glucose yields from the selected models of each dataset are shown. As the axes in the plots are defined by the predicted values, the higher the linearity of scatter points the higher the correlation between two models. The covariance can be statistically described by the Pearson coefficients shown in the upper part of Fig. 3. For better visualization, higher correlations are presented in darker colors. In both datasets, ppr and cubist methods have the highest linearities with Pearson coefficient over 0.8 compared to other pairs. The ppr method is a nonlinear transformation of inputs added in linear combinations similar to neural networks. As cubist is a tree-based regression model consisting of linear models at the node of the tree, the results showed that the projection from ppr had higher covariance compared to cubist after it performed the nonparametric regression to our datasets. However, they were selected due to the low RMSE so the accuracy of the ensemble model can be ensured.

3.3. Ensemble model

Previous studies showed that introducing randomness and diversity by using models of different topologies, manipulating features, and changing training subsets can improve the generalization in the aggregation of models running independently of each other [25–27]. A two-fold ensemble model stacked with base models was developed in order to improve the final prediction by integrating models to lower the bias from the predictions of the first layer. Instead of the original features from the dataset, the predictions from base models were used as the inputs in the second layer model, and the parameters were evaluated along with the previous predictions in the algorithms adopted in the ensemble model. As other studies showed [28,29], ensemble methods can be categorized as linear and nonlinear methods. In this study, a linear weighted ensemble and a gbm model were used in the second layer by combining the learning outputs in a linear-fashioned and a nonlinear stacked model by gradient boosting respectively.

The results shown in Tables 2 and 3exhibit the RMSE, weighted values in linear ensemble model, and the relative influence from base learners in gbm ensemble for both datasets. Reduction of RMSE in nonlinear gbm ensemble was -19.88 % and -20.59 % compared to linear ensemble model, which shows that the nonlinear ensemble model outperforms the linear-weighted ensemble model in both datasets, with close to 20 % reduction in RMSE. As shown in Table 2, the linear weighted value and the relative influence for each base model have different significance in the linear and nonlinear ensemble models, implying the different ways of evaluation observed from base learners in linear and nonlinear models as described in Section 2.3. In dataset 1, bridge had the most influence in both ensemble models while ppr and rpart had the least influence in linear and nonlinear ensembles respectively. In dataset 2, rf and cubist had the least contribution to linear weighted and gbm models respectively, while xgbTree had the highest impact to both ensemble models. The level of contribution did not correspond to the RMSE from a single model. For example, ppr had the lowest RMSE compared with other models in dataset 1, while its weighted value was the least in linear-weighted ensemble model. The results reflect that the ensemble models considered not only the accuracy between predicted and actual value, but also took into account the sensitivity from fluctuations in training data. To further improve the prediction performance and prevent overfitting, filtering out features with little influence to glucose yield was conducted and examined.

3.4. Feature selection

By using an extreme gradient boosting (xgbLinear) model to analyze the feature importance, the influence from each variable was evaluated and presented in Fig. 4. The mean squared error of regression was evaluated when the variable is used for splitting at each separation of the tree, and the improvement of the mean squared error was calculated. The improvement from each variable in the trees was averaged and normalized, and used to determine the importance of the variable [30]. Features above the level of importance 20 were selected in model building, where pretreatment temperature and time, pretreatment method (IL), and the ratios of SG and EC were considered in dataset 1, and pretreatment method (IL), pretreatment and time, and CS ratio were considered in dataset 2. After selecting the desired variables, the data

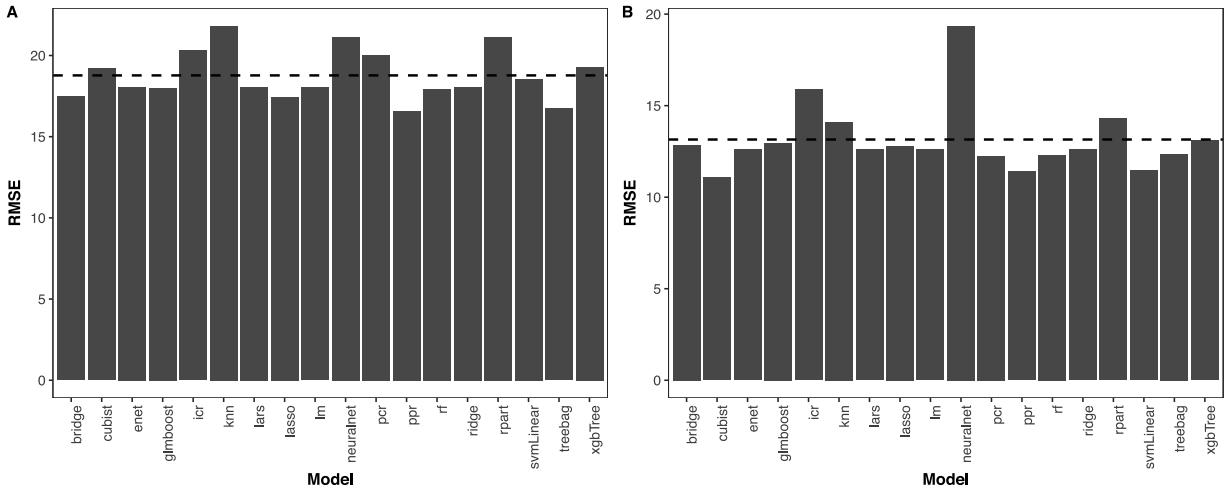

Fig. 2. RMSE from eighteen individual models for (A) dataset 1 and (B) dataset 2. The mean of the RMSE was labeled by the dashed line.

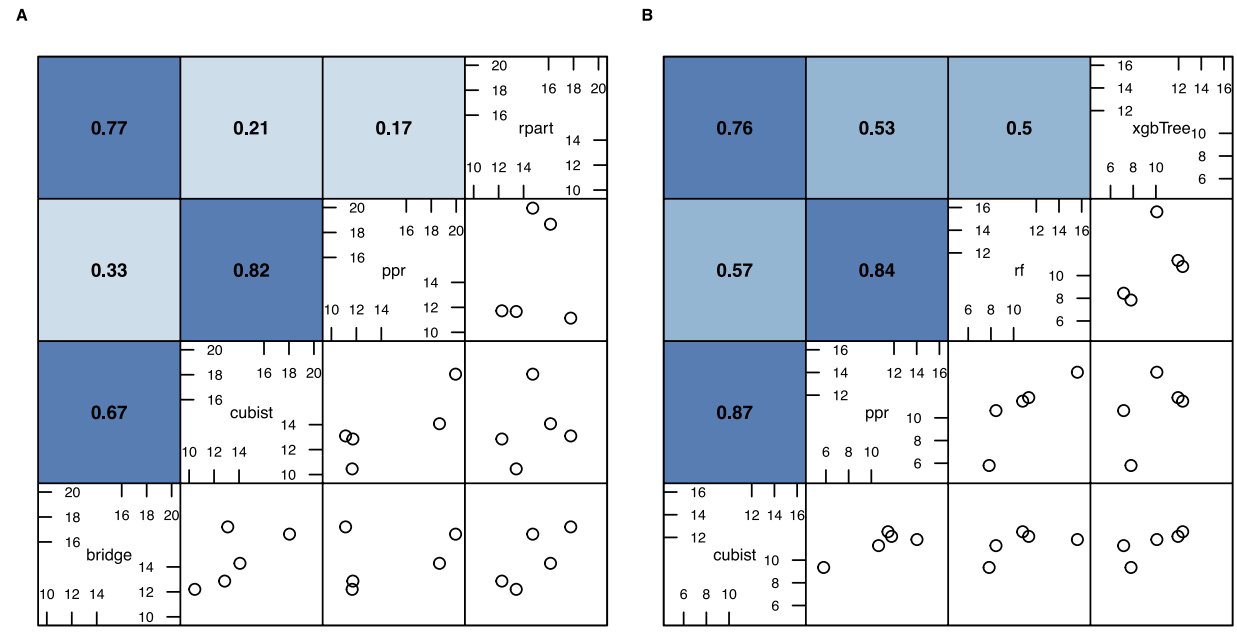

Fig. 3. Correlations (Pearson correlation coefficients) between MAE among first layer models for (A) dataset 1 and (B) dataset 2. The correlation was reflected by the color gradient in the upper part of the figure and the linearity of the points in the bottom part.

Table 2
Summary and comparison of ensemble model performance in two datasets.

Dataset 1				
Model	bridge	cubist	ppr	rpart
Linear-weighted*	−1.60	0.99	0.13	1.14
RMSE (test set)	24.40			
GBM: Relative influence	35.13	31.33	30.05	3.49
RMSE (test set)	19.55			
Dataset 2				
Model	cubist	ppr	xgbTree	rf
Linear-weighted*	−0.54	0.69	0.74	−0.014
RMSE (test set)	14.86			
GBM: Relative influence	2.95	31.27	52.73	13.05
RMSE (test set)	11.80			

* Intercept equals to 24.54 and 8.29, respectively.

Table 3
Summary and comparison of ensemble model performance in two datasets after feature selection.

Dataset 1				
Model	bridge	cubist	ppr	rpart
Linear-weighted*	0.65	0.63	0.22	−0.30
RMSE (test set)	18.35			
GBM: Relative influence	18.69	15.83	60.78	4.69
RMSE (test set)	16.96			
Dataset 2				
Model	cubist	ppr	xgbTree	rf
Linear-weighted*	−0.57	0.69	0.55	0.34
RMSE (test set)	12.71			
GBM: Relative influence	14.41	21.77	51.38	12.44
RMSE (test set)	9.35			

* Intercept equals to -12.90 and 0.37 respectively.

with only chosen features were used to build two-fold ensemble models as described in Section 3.2. The relative influences from base models were similar as before, but the changes in RMSE for both linear and nonlinear ensembles in two datasets were -24.80 % and -13.25 % compared to ensembles using all features in dataset 1, and -7.67 % and -20.42 % in dataset 2, see Table 3. Comparing the prediction RMSE of non-linear gbm ensemble to linear ensemble model, the RMSE was reduced -7.57 % and -26.44 % for two datasets respectively. It should be noted that initially, only two features, temperature and pretreatment method (IL) with importance over 50 were considered for dataset 1 and dataset 2. Due to the small size of data available to us, a large reduction of the number of features in the models can potentially lead to an unbalance in data and thereby bias. Although the results were not shown, RMSE increased significantly when a high threshold was applied for both datasets. Further, the importance threshold of 50 was preventing us from examining the effect from individual feedstocks in the blends. Adjusting the importance threshold allowed for feedstocks to be included without other features that carry negligible importance. This helped the model to prevent overfit from unrelated features, while studying the effect from blended feedstock. The positive results in RMSE from testing and validation data has proven the effectiveness of our design methodology. This helps the model to prevent overfit from unrelated features, while the effect from blended feedstock is still included. After lowering the accepted level of importance from 50 to 20, the RMSEs from both linear and nonlinear ensemble models showed improvement for both datasets. Although the RMSEs after feature selection were lower, the RMSE for the linear weighted ensemble improved much more than nonlinear ensemble in dataset 1 while the opposite was true in dataset 2.

Per the error density plot in Fig. 5, the ensemble models after feature selection had smoothened the variance and lowered the bias as shown from the changes in peak shape. As generally simpler models (e.g. linear regression) have higher bias but lower variance while more complex ones (e.g. decision trees) have lower bias and are more sensitive to the changes in datasets, the results from the ensembles had better tradeoff in balancing bias and variance in a fixed size dataset (< 100 data points in this study) than a single model especially after feature selection. However, it was difficult to ascertain whether the model with the least RMSE is the best analysis method for future prediction as it may have simply overfitted to the training data. Testing the models with validation data is essential to ascertain if the models can explain the biological phenomena accurately.

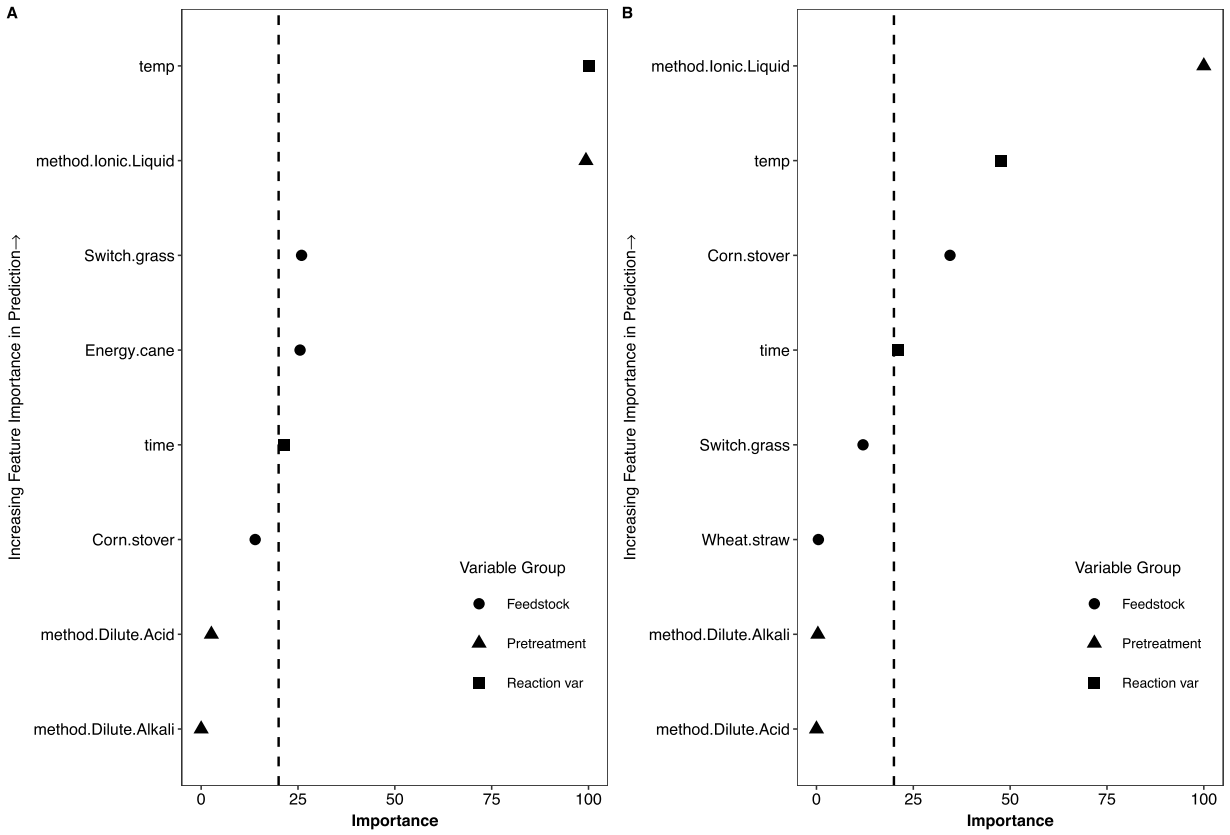

Fig. 4. Relative influence from process variables of pretreatment catalysts (Pretreatment), feedstock blends (Feedstock), and reaction variables (Reaction var) in (A) dataset 1 and (B) dataset 2. The selection threshold of the level of importance is represented by the dashed line.

3.5. Model validation

3.5.1. Dataset 1

In our previous studies, to verify our LRM, separate experiments were conducted to generate validation data [9]. To verify the goodness of fit for models constructed in this study, the same validation data were applied to the base learners and ensemble models and the results are shown in Fig. 6. The RMSE from the validation dataset was lower by 30 %–60 % compared to that from training data when applied to same ensemble models presented in Table 4, which confirmed no significant overfitting in our model training process. When all variables were applied as features, the nonlinear gbm ensemble model performed the best. Linear stacked ensemble outperformed other models after feature selection in dataset 1. As discussed in Section 3.3., the ensemble model balances the performance from base learners. Since it is more neutral in bias-variance tradeoff, the linear-weighted ensemble model showed better fitting from low (40 %) to high glucose yield (90 %) compared to individual models. While single models may predict glucose yield better when the value of it is lower than 50 %, the performance in other ranges may be poorer, leading to bigger errors compared to ensemble models.

Fig. 6(A) shows the R^2 and MAE for data in different models, where linear-weighted ensemble and the linear mixed model from JMP (LRM) share similar R^2 and MAE, but linear weighted ensemble was preferable overall. Regardless of their type, both linear or nonlinear ensemble models have demonstrated better prediction performance in RMSE than single models before and after feature selection. Both the regression coefficients and RMSE prove that when an appropriate model or data filtration has been applied, reliable prediction on glucose yield from an ensemble model can be achieved by applying values of feedstock blend ratios and other experimental variables.

3.5.2. Dataset 2

Following the same procedure as for dataset 1, 16 data points outside of the training and testing dataset were used for validation for dataset 2. Both LRM and ML models developed in R were included. The RMSE was similar but not lower compared to test data for both ensembles and single models as shown in Table 4. For dataset 2, ppr model performed the best without feature selection while gbm ensemble outperformed other models after features with negligible influence were filtered out. The lowest glucose yield in the validation data was at 14 %, which is out of the glucose range in the training dataset and may be the main factor that interferes with the total RMSE. From Fig. 6(B), all three models (ppr, nonlinear ensemble, and LRM) showed better predictions when the glucose yield was over 50 %, which was the majority representation in the training dataset. Although in Section 3.3 the ensembles showed lesser bias and variance both for dataset 1 and dataset 2, the result in validation dataset 2 revealed that data outside of the training dataset may lead to error in prediction. The R^2 and MAE values shown in Fig. 6 (B) also had poorer performance compared to dataset 1. However, ensemble models became the best performers after feature selection, which again bolsters the previous observation that feature selection can improve the performance in bias and variance for ensemble over single models. The RMSEs from previously developed LRM was 8.01 for dataset 1 and 14.65 for dataset 2 (Table 4). Comparing the results with ML models, for both the datasets, only the ratios of feedstocks in the blends had a significant impact on glucose yields (p-value < 0.0001) in LRM. While in the ensemble models, the influence from the reaction variables and pretreatment catalysts was included by implementing a decision tree in feature selection to achieve better predictions.

The goal of the deconstruction (pretreatment and saccharification) is to achieve the highest glucose yield from feedstock blends. The experiments were designed such that we had a large range of glucose yields and thereby we could successfully build a robust model to obtain

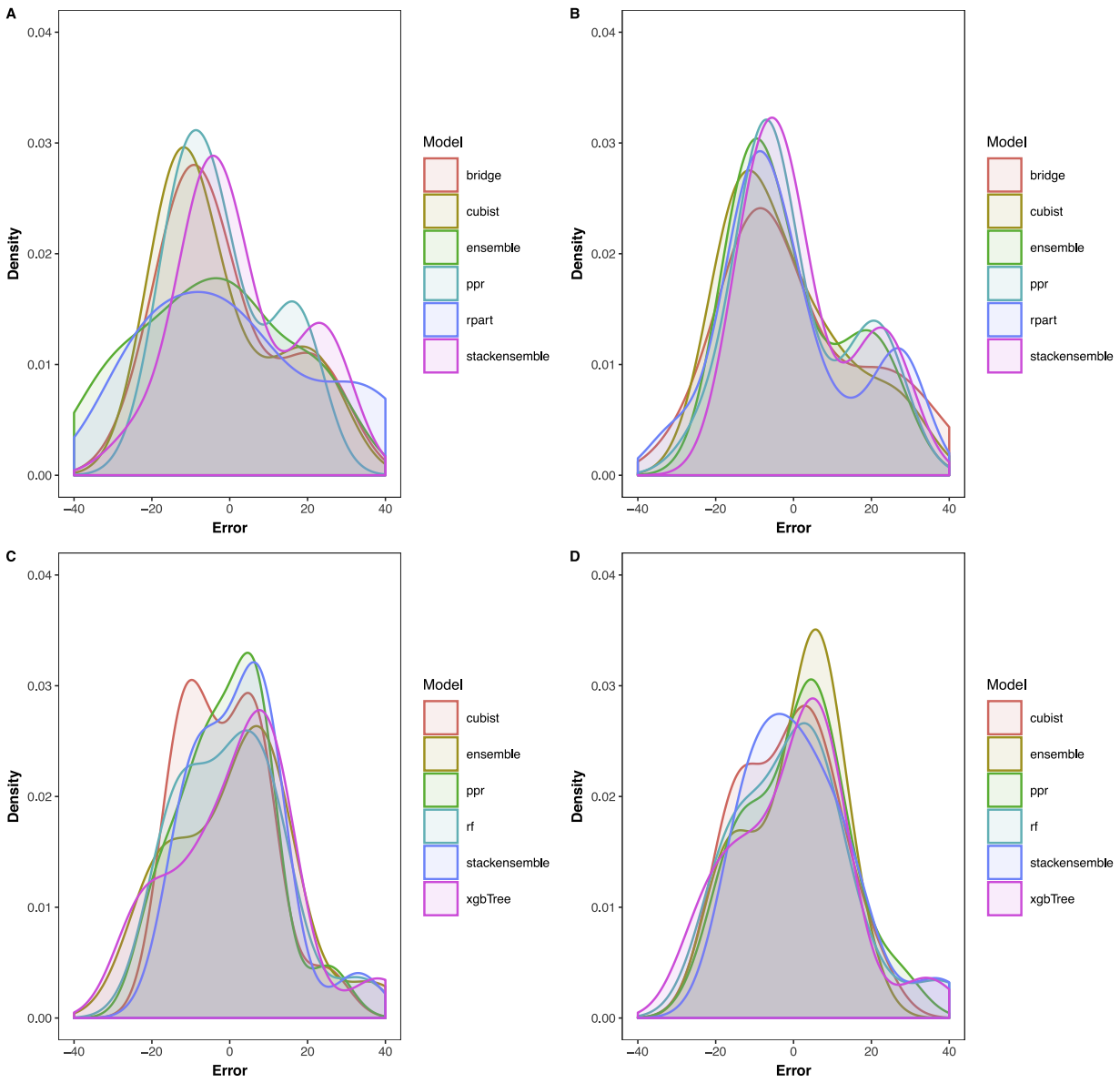

Fig. 5. Density plots of the prediction errors from base learners and ensemble models. Base learners were named by model names, and linear / nonlinear ensembles were represented as "ensemble" and "stackensemble" respectively. (A) and (B) are data from dataset 1 with all variables as features and selected features respectively. (C) and (D) are data from dataset 2 with all variables as features and selected features respectively.

accurate predictions at either end of the range. The linear-weighted model and LRM from dataset 1 however, converged more at higher glucose yields, typically over 80 %. The same results appeared in the nonlinear ensemble model and LRM in dataset 2. As such, we have to perform more experimental studies, especially varying our pretreatment process conditions, to obtain lower glucose yields and thereby training data that can enable the models to predict better in the lower range.

3.6. Perspective

We selected strong and diverse models as the base layer and directed their predictions inputs to the second layer to generate the final predictions from the ensemble model to offset the weakness and bias from individuals. The feature analysis improved our ability to predict glucose yield from many feedstock blends treated with various pretreatment chemistries at different temperatures. From Fig. 4, IL pretreatment was observed to be most effective in converting feedstock blends to glucose, and reaction temperature had a stronger impact than reaction time in our studies. Base on the pretreatment data, ML models calculated optimal feedstock blends while maximizing low cost, locally available feedstock concentration and glucose yields. Maximizing the blending of local feedstocks can help ensure that bio-based manufacturing is possible in states like Florida and Kansas. Although from dataset 2, we see a higher dependency on CS when blended with local WS, local SG and EG have higher impact to glucose yield in dataset 1. Blending feedstocks, nonetheless, has the potential to relieve some stress on the biomass supply chain.

From dataset 2, we also identified the optimal deconstruction conditions to produce at least 45 % (of theoretical) glucose yields, by varying pretreatment catalysts and reaction temperature and time. The best feedstock blends along with reaction parameters for both datasets are described in Table 5. For ddataset 1, both blends can demonstrate glucose yield over 90 %, with reduced ratio in CS by blending feedstocks with SG and EC. One combination can even decrease CS to only 0.07 fraction in the blending portion with less temperature and longer time. For dataset 2, it is possible to have glucose yield over 80–90 % with only WS as the feedstock. Another blend with 0.33 CS and 0.34 SG provided similar glucose yield with similar reaction conditions. The results

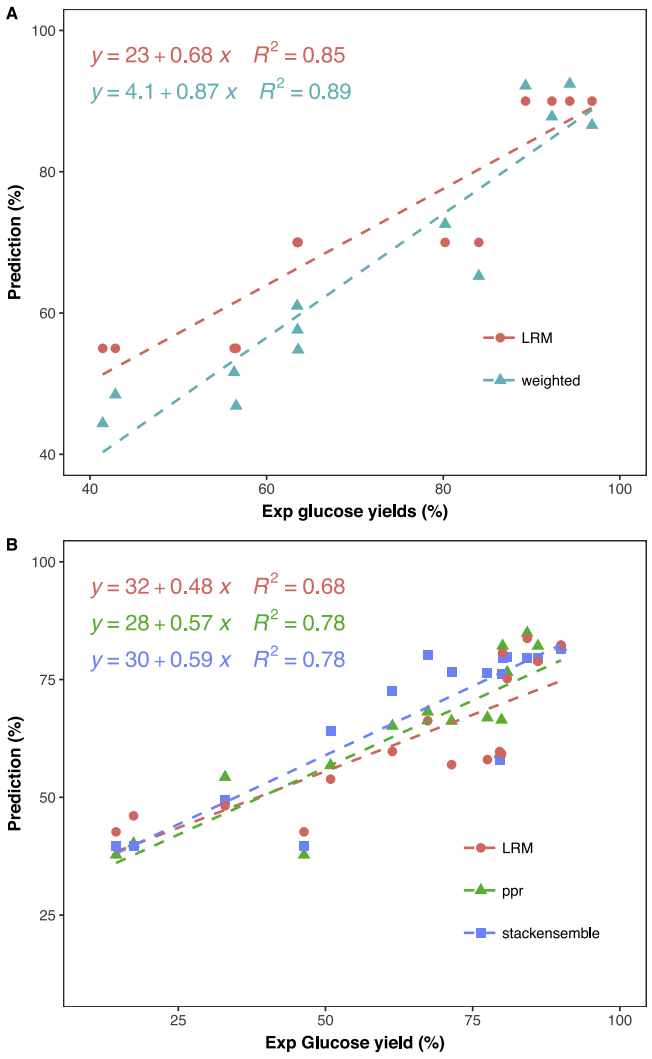

Fig. 6. Linear correlation (dashed line) for modeling on glucose yield from (A) dataset 1: Linear ensemble (weighted), and linear regression model (LRM) developed in JMP. R^2 are 0.89 and 0.85 respectively. MAEs are 6.61 and 6.65 respectively. (B) dataset 2: Linear correlation (dashed line) for modeling on glucose yield from ppr, gbm ensemble (stackensemble), and LRM. R^2 are 0.78, 0.78, and 0.68 by order. MAEs are 9.70, 10.06, and 11.11 respectively. Linear equations were colored based on models.

demonstrate that similar glucose yield can be achieved with different feedstock blends, and it is predictable by our models. Our models can provide such calucations in real-time, offering solutions for feedstock supply chain issues in the biorefinery industry. It should be emphasized that the LRM model described in Narani et al. [9] was developed to predict continuous envelopes of biomass blends that are optimal for a given pretreatment condition to achieve a predetermined sugar yield or vice versa (Fig. S4). ML models in the study were designed with the same concept but to extend beyond the envelope. In this study, we also predicted continuous envelops of optimal process parameters and examined impact from features Fig. 4. The models were designed to be flexible depending on the conditions, and, when applied, such predictive models will reduce biorefineries' dependence on singular feedstocks. Such predictions can also be very useful to a biorefinery managing process conditions as they experience feedstock variability on a day-to-day basis.

The nonlinear ensembles and linear-weighted ensembles led to reliable predictions in both dataset 1 and 2. Training errors that occurred from finite-size sampling were addressed by feature selection and ensemble technique. Feature selection can help improve model performance for the ensemble ones while the impact is not obvious for individual models. By choosing an appropriate filter area for the features that includes the blends but suppressing the variants that have lower influence in predictions, we were able to reduce RMSE. There are other ML techniques that can be further applied, including filtering out outliers, tuning more hyperparameters in each model or building more folds of the ensembles. However, the more sophisticated the ensemble is, the higher the possibility of losing balance between bias-variance [31–33]. Finding the compromise between underfitting and overfitting is a common dilemma in designing predictive models by machine learning.

By the nature of the small datasets in this study (< 100 data points), the complexity of the ensemble model was controlled with four base models and two layers to prevent overfitting. It should be noted that inappropriate filtering can lead to underfitting or overoptimization in the final results. To prevent the generation of an error-prone model, outliers in the dataset were checked beforehand and models were constructed in a relatively conservative manner - the choice of number of models and the threshold level in feature selection was kept to a minimum. We also ensured the models passed proper testing and validation with similar or better RMSE compared to training data so the models do not under / overfit. The training dataset should cover a wide range of glucose yield for an ensemble to be able to capture more comprehensive information. In both linear and nonlinear models, data size can significantly affect the final influence in the linear slope or the split points in trees. The bigger the data set, the more un-correlated aspects can be collected by different base models. Since our datasets do not fall in the range of big data and the inter-correlation between models were not less than 0.5, 3-fold ensembles or more blends in different levels and more tuning were not added to the final model. The prediction power in this study can be guaranteed only in the range of the training dataset. Increasing the size of the dataset and covering a more comprehensive range in the features should greatly help us in developing a predictive model that is applicable to a wide range of feedstock blends and process conditions leading to a wide range of glucose yields. The examination of process features can direct researchers focus towards studying the most impactful variables, such as reaction temperature or the type of feedstocks. It may be beneficial to construct a feedstock library for the models and expand to the feedstocks available across the nation's geographic expanse. Future work geared towards generating and adding more data to this and other such approaches can help in designing a

Table 4
Lineup of RMSE of validation data among various models for two datasets.

Dataset 1							
Model	Weighted	GBM stack	bridge	cubist	ppr	rpart	LRM
RMSE (full feature)	16.50	9.08	11.39	9.50	10.03	18.90	8.01
RMSE (feature selection)	7.95	9.87	11.24	9.42	9.56	12.90	8.01
Dataset 2							
Model	Weighted	GBM stack	cubist	ppr	xgbTree	rf	LRM
RMSE (full feature)	15.29	14.10	13.35	12.46	15.38	14.31	14.65
RMSE (feature selection)	14.10	12.60	14.47	14.46	14.10	14.77	14.65

Table 5
Feedstock blends with the maximum glucose yield for two datasets from the models.

Dataset 1								
Pretreatment	CS	SG	EC	Temp (C)	Time (min)	Glucose yield (%)	LRM	Linear weighted
IL	0.24	0.22	0.54	151.92	102.67	94.33	90	92.42
IL	0.07	0.05	0.88	131.72	180	96.85	90	86.59

Dataset 2									
Pretreatment	CS	SG	WS	Temp (C)	Time (min)	Glucose yield (%)	LRM	stack ensemble	ppr
IL	0	0	1	120	180	90	82.34	81.45	81.95
IL	0.33	0.34	0.33	129.6	180	84.26	83.75	79.58	84.91

more robust and accurate predictive model that can unravel the effect from each process parameter to overall bio-based production.

4. Conclusions

ML models have been sparsely applied to understand and predict glucose yields from biomass deconstruction. In this study, an ensemble model was stacked by selecting base learners after screening individual models from a wide range of regression types. With additional feature selection, we improved model performance by 20 %. As both linear and nonlinear ensemble models exhibited good prediction capability, we concluded that ensemble combined with feature selection will provide best predictions. ML ensemble models are agile models that can help predict glucose yields from a wide range of feedstocks and process conditions, and give rich insights into predicting data unlike those from linear mixed models. The models can maximize sugar yields from the dynamic blending space and assist later bioprocesses such as fermentation in the biomanufacturing. Expanding the quality and quantity of data used to train these models can help apply them in a biorefinery setting where real-time decisions are necessary to process variable feedstocks.

Credit author statement

This work was funded by US Department of Energy through Bio-Energy Technologies Office (BETO) under an Annual Operating Plan.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank researchers from Idaho National Laboratory: Allison Ray, Chenlin Li, and Damon Hartley for previous collaborations that helped generate data included in this publication. We would also like to thank other Lawrence Berkeley collaborators: NVSN Murthy Konda, Phil Coffman, James Gardner, Todd R Pray, and Blake Simmons for their previous contributions that made this work possible.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.bej.2020.107896.

References

[1] J.R. Regalbuto, Cellulosic biofuels–got gasoline? Science 325 (2009) 822–824, https://doi.org/10.1126/science.1174581, 80.
[2] Y. Zhang, P. Bi, J. Wang, P. Jiang, X. Wu, H. Xue, J. Liu, X. Zhou, Q. Li, Production of jet and diesel biofuels from renewable lignocellulosic biomass, Appl. Energy 150 (2015) 128–137, https://doi.org/10.1016/j.apenergy.2015.04.023.
[3] M.H. Langholtz, B.J. Stokes, L.M. Eaton, 2016 Billion-ton Report: Advancing Domestic Resources for a Thriving Bioeconomy, Volume 1: Economic Availability of Feedstock, Oak Ridge Natl. Lab. Oak Ridge, Tennessee, Manag. By UT-Battelle, LLC US Dep. Energy, 2016, 2016, pp. 1–411.
[4] J. Sadhukhan, K.S. Ng, E.M. Hernandez, Biorefineries and Chemical Processes: Design, Integration and Sustainability Analysis, John Wiley & Sons, 2014.
[5] A. Aden, M. Ruth, K. Ibsen, J. Jechura, K. Neeves, J. Sheehan, B. Wallace, L. Montague, A. Slayton, J. Lukas, Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-Current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis for Corn Stover, 2002 (accessed June 2, 2019), https://apps.dtic.mil/docs/citations/ADA436469.
[6] A.M. da Costa Lopes, K.G. João, D.F. Rubik, E. Bogel-Łukasik, L.C. Duarte, J. Andreaus, R. Bogel-Łukasik, Pre-treatment of lignocellulosic biomass using ionic liquids: wheat straw fractionation, Bioresour. Technol. 142 (2013) 198–208, https://doi.org/10.1016/J.BIORTECH.2013.05.032.
[7] S. Shields, R. Boopathy, Ethanol production from lignocellulosic biomass of energy cane, Int. Biodeterior. Biodegradation 65 (2011) 142–146, https://doi.org/10.1016/J.IBIOD.2010.10.006.
[8] C. Li, B. Knierim, C. Manisseri, R. Arora, H.V. Scheller, M. Auer, K.P. Vogel, B. A. Simmons, S. Singh, Comparison of dilute acid and ionic liquid pretreatment of switchgrass: biomass recalcitrance, delignification and enzymatic saccharification, Bioresour. Technol. 101 (2010) 4900–4906, https://doi.org/10.1016/J.BIORTECH.2009.10.066.
[9] A. Narani, P. Coffman, J. Gardner, C. Li, A.E. Ray, D.S. Hartley, A. Stettler, N.V.S.N. M. Konda, B. Simmons, T.R. Pray, D. Tanjore, Predictive modeling to de-risk bio-based manufacturing by adapting to variability in lignocellulosic biomass supply, Bioresour. Technol. (2017), https://doi.org/10.1016/j.biortech.2017.06.156.
[10] A. Narani, N.V.S.N.M. Konda, C.S. Chen, F. Tachea, P. Coffman, J. Gardner, C. Li, A.E. Ray, D.S. Hartley, B. Simmons, T.R. Pray, D. Tanjore, Simultaneous application of predictive model and least cost formulation can substantially benefit biorefineries outside Corn Belt in United States: a case study in Florida, Bioresour. Technol. (2019), https://doi.org/10.1016/j.biortech.2018.09.103.
[11] V.B. Agbor, N. Cicek, R. Sparling, A. Berlin, D.B. Levin, Biomass pretreatment: fundamentals toward application, Biotechnol. Adv. 29 (2011) 675–685, https://doi.org/10.1016/J.BIOTECHADV.2011.05.005.
[12] P. Kumar, D.M. Barrett, M.J. Delwiche, P. Stroeve, Methods for pretreatment of lignocellulosic biomass for efficient hydrolysis and biofuel production, Ind. Eng. Chem. Res. 48 (2009) 3713–3729, https://doi.org/10.1021/ie801542g.
[13] N. Mosier, C. Wyman, B. Dale, R. Elander, Y.Y. Lee, M. Holtzapple, M. Ladisch, Features of promising technologies for pretreatment of lignocellulosic biomass, Bioresour. Technol. 96 (2005) 673–686, https://doi.org/10.1016/J.BIORTECH.2004.06.025.
[14] Y. Zheng, Z. Pan, R. Zhang, Overview of biomass pretreatment for cellulosic ethanol production, Int. J. Agric. Biol. Eng 2 (2009) 51–68, https://doi.org/10.25165/IJABE.V2I3.168.
[15] J.A. Cruz, D.S. Wishart, Applications of machine learning in Cancer prediction and prognosis, Cancer Inform. 2 (2006), https://doi.org/10.1177/117693510600200030, 117693510600200.
[16] P. Langley, H.A. Simon, Applications of machine learning and rule induction, Commun. ACM 38 (1995) 54–64, https://doi.org/10.1145/219717.219768.
[17] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (2015) 321–332, https://doi.org/10.1038/nrg3920.
[18] J. Fischer, V.S. Lopes, S.L. Cardoso, U. Coutinho Filho, V.L. Cardoso, Machine learning techniques applied to lignocellulosic ethanol in simultaneous hydrolysis and fermentation, Braz. J. Chem. Eng. 34 (2017) 53–63, https://doi.org/10.1590/0104-6632.20170341s20150475.
[19] F. Xu, Z.-W. Wang, Y. Li, Predicting the methane yield of lignocellulosic biomass in mesophilic solid-state anaerobic digestion based on feedstock characteristics and process parameters, Bioresour. Technol. 173 (2014) 168–176, https://doi.org/10.1016/J.BIORTECH.2014.09.090.
[20] D.L.B. Fortela, W.W. Sharp, E.D. Revellame, R. Hernandez, D. Gang, M.E. Zappi, Computational evaluation for effects of feedstock variations on the sensitivities of biochemical mechanism parameters in anaerobic digestion kinetic models, Biochem. Eng. J. (2019), https://doi.org/10.1016/j.bej.2019.01.001.
[21] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232. http://www.jstor.org/stable/2699986.
[22] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.

[23] B. Leo, Bagging predictors, Mach. Learn. 24 (1996) 123–140 (accessed June 4, 2019), https://link.springer.com/content/pdf/10.1007/BF00058655.pdf.
[24] J.A. Hartigan, Printer graphics for clustering, J. Stat. Comput. Simul. (1975), https://doi.org/10.1080/00949657508810123.
[25] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, Adv. Neural Inf. Process. Syst. 7 (1995) https://doi.org/10.1.1.37.8876.
[26] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. (2003), https://doi.org/10.1023/A:1022859003006.
[27] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. (1999), https://doi.org/10.1613/jair.614.
[28] S.S. Rathore, S. Kumar, Linear and non-linear heterogeneous ensemble methods to predict the number of faults in software systems, Knowledge-Based Syst. 119 (2017) 232–256, https://doi.org/10.1016/J.KNOSYS.2016.12.017.
[29] J. Sill, G. Takacs, L. Mackey, D. Lin, Feature-Weighted Linear Stacking, ArXiv Prepr, 2009. ArXiv0911.0460 http://arxiv.org/abs/0911.0460 (accessed May 29, 2019).
[30] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (2013) 21, https://doi.org/10.3389/fnbot.2013.00021.
[31] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the Bias/Variance dilemma, Neural Comput. 4 (1992) 1–58, https://doi.org/10.1162/neco.1992.4.1.1.
[32] M. Keijzer, V. Babovic, Genetic Programming, Ensemble Methods and the Bias/Variance Tradeoff – Introductory Investigations, Springer, Berlin, Heidelberg, 2000, pp. 76–90, https://doi.org/10.1007/978-3-540-46239-2_6.
[33] P. Munro, H. Toivonen, G.I. Webb, W. Buntine, P. Orbanz, Y.W. Teh, P. Poupart, C. Sammut, C. Sammut, H. Blockeel, D. Rajnarayan, D. Wolpert, W. Gerstner, C. D. Page, S. Natarajan, G. Hinton, Bias variance decomposition. Encycl. Mach. Learn, Springer US, Boston, MA, 2011, pp. 100–101, https://doi.org/10.1007/978-0-387-30164-8_74.