Machine learning models for the prediction on efficacy of ionic liquid-aided biomass pretreatment

Biswanath Mahanty  
Karunya Institute of Technology and Sciences

Munmun Gharami  
Karunya Institute of Technology and Sciences

Dibyajyoti Haldar  
dhaldar@karunya.edu
Karunya Institute of Technology and Sciences

Research Article

Keywords: Artificial neural network, Bayesian optimization, Hyperparameters, Permuted predictor importance, Predictor selection, Random Forest

Posted Date: February 6th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3916542/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at BioEnergy Research on March 26th, 2024. See the published version at https://doi.org/10.1007/s12155-024-10747-2.
Abstract

The influence of ionic liquids (ILs) characteristics, lignocellulosic biomass (LCB) properties, and process conditions on LCB pretreatment is not well understood. In this study, a total of 129 experimental data on cellulose, hemicellulose, lignin, and solid recovery from IL-based LCB pretreatment were compiled from literature to develop machine learning models. Following data imputation, bilayer artificial neural network (ANN) and random forest (RF) regression were developed to model the dataset. The full-featured ANN following Bayesian hyperparameter (HP) optimization though offered excellent fit on training (R^2=0.936–0.994), cross-validation (R^2CV) performance remained marginally poor, i.e., between 0.547 and 0.761. The fitness of HP-optimized RF models varied between 0.824–0.939 for regression, and between 0.383–0.831 in cross-validation. Temperature and pretreatment time had been the most important predictors, except for hemicellulose recovery. Bayesian predictor selection combined with HPO improved the R^2CV boundary for ANN (0.555–0.825), as well as for RF models (0.474–0.824). As the predictive performance of the models varied depending on the target response, the use of a larger homogeneous dataset may be warranted.

1. Introduction

The exploration of renewable energy resources is an integral part of the UN's sustainable development goals SDG 7, which aims not only to reduce dependency on the finite reserve of fossil fuels but also to minimise the environmental impact from the release of harmful greenhouse gases (Li et al., 2022; Su et al., 2023). In recent times, lignocellulosic biomass (LCB) has received much appreciation as a renewable source of energy, as just 8% of annual LCB production can compensate for the yearly consumption of fossil fuels (Stark, 2011). Though easily and widely available in nature, the proportion of cellulose, hemicellulose and lignin varies with the biomass type, geographical location or season (for the same biomass) (Luo et al., 2022; Maibam & Goyal, 2022). The intricate association of carbohydrate fractions (cellulose, hemicellulose) with lignin, maintained through hydrogen and covalent bonds (Shabbirahmed et al., 2023), must be disrupted before conversion of LCB into usable biofuels like bioethanol, biodiesel or biogas.

Different physical (Ding et al., 2022), chemical (Y. Chen et al., 2020), physico-chemical (Sarker et al., 2021) and biological (Guo et al., 2022) pretreatment methods are employed to increase accessibility of cellulose and hemicellulose fractions. In recent years, ionic liquids (ILs) have emerged as one of the environment-friendly organic solvents with several impressive properties like non-flammability, low conductivity and high stability. ILs are salts of organic cationic and anionic species with delocalized charge distribution having melting points lower than the boiling point of water (Alayoubi et al., 2020; Brandt-Talbot et al., 2017). Functionalyzed cations or anions in ILs offer greater specificity in different chemical reactions. Use of ILs, such as [Emim][OAc] (1-ethyl-3-methylimidazolium acetate), [Bmim][Cl] (1-buty-3-methylimidazolium chloride), [Hmim][Cl] (1-hexyl-3-methylimidazolium chloride) and [Omin][Cl] (1-octyl-3-methylimidazolium chloride) has already been demonstrated in pretreatment of both hardwood and softwood biomass at moderate temperature (Alayoubi et al., 2020; Brandt-Talbot et al., 2017; Dotsenko et al., 2018). Apart from reaction conditions, or solid loading, the biomass pretreatment efficiency can vary depending on the nature of ILs as well as biomass properties (Z. Chen et al., 2022; Chourasia et al., 2021). An exhaustive exploration of such a large variable space for the identification of influential factors and their optimal setting through laboratory experiments is extremely tedious, if not impossible.

In recent times, a variety of machine learning (ML) tools viz., artificial neural network (ANN), random forest (RF), support vector regression (SVR), etc. and their combinations are being adopted to model the influence of process conditions, biomass properties on pretreatment efficiency (Aruwajoye et al., 2021; Smuga-Kogut et al., 2021; Xu et al., 2023). ANN models are extremely competent in mapping the non-linear relationships among the process parameters and responses in inorganic salt-based (Moodley et al., 2019), enzymatic (Vani et al., 2015), acid-enzymatic (Haldar et al., 2023), and IL-based LCB treatment (Smuga-Kogut et al., 2021). RF is an ensemble algorithm where predictions from many regression tree models built on bootstrap data are averaged, providing good accuracy (Torres-Barrán et al., 2019).

Though ML-based modelling approaches are gaining significant interest, overfitting, and poor generalization capabilities could be a concern, particularly with a limited dataset size, presence of uninformative predictors, or unnecessarily complex ML architecture (Tian & Zhang, 2022). Researchers are adopting ambitious large-scale (>100) batch experiments to fulfil the data requirements for ML-based modelling (Smuga-Kogut et al., 2021). Curation of experimental data from existing literature, though rarely explored, could be a faster and more economical alternative. However, such a compilation would certainly have missing values, as the dimension of predictors or even responses is likely to vary across studies (Emmanuel et al., 2021), and appropriate imputation techniques need to be adopted (Wang et al., 2022).

The generalization quality of ML can be improved by retaining informative variables (Kim & Kim, 2019), but no explicit procedure do exist for the same. Predictor importance matrix ranks the variables based on their significance, but does perform model refitting excluding those unimportant variable(s) or their combinations (Chehreh Chelgani et al., 2016). Similarly, the configurable hyperparameters (HPs) for ML algorithms defining the architecture must be fine-tuned to achieve stability, speed and performance (Yang & Shami, 2020). Though many HP optimization (HPO) algorithms have been developed over the last decades, they are either not widely used, or used with incorrect search space (Bischl et al., 2023). The use of incorrect fitness indices, e.g., regression coefficient (R^2) rather than cross-validation R^2 in training and/or ML model comparison could be an additional setback in many studies (Kanthasamy et al., 2023).

In this work, experimental data on IL-based treatment of biomass focusing on physical properties of IL, biomass composition, and treatment conditions was gathered from literature. Three different data imputation techniques, i.e., k-nearest neighbours (ANN), principal component analysis (PCA), and RF regression, were compared to address the missing values. ANN, and RF models were developed for each of the responses, separately. Following assertion of boundaries for number of neurons in hidden layer in ANN models through grid search, Bayesian optimization was adopted to fine-tune all HPs. OOB predictor importance from full-predictor HP optimized RF model served as guide for forward selection and backward elimination of predictors. Results of Both ANN, and RF were compared for their suitability in dealing with heterogeneous data for IL-based biomass pretreatment.
2. Materials and Methods

2.1 Data reconciliation and preprocessing

2.1.1. Data collection

Information on IL-based pretreatment of biomass from existing studies was collected through a manual search in literature published by Elsevier, Springer, ACS (American Chemical Society) and RSC (Royal Society of Chemistry) in the last 10 years (2014–2023). The information gathered for a total of 129 data was on properties of IL (i.e., molecular weight, density, viscosity, conductivity, and melting point), characteristics of biomass (i.e., particle size, content of cellulose, hemicellulose, lignin, and other extractive), pretreatment conditions (reaction temperature, time, and solid/liquid ratio). These 13 independent variables are known for their perceived importance in IL pretreatment of different kinds of lignocellulosic residues. Four different responses (i.e., solid recovery, cellulose, hemicellulose, and lignin recovery) from IL-based pretreatment were considered in the data set as defined in Eq. 1–2:

\[ SR = \left( \frac{W_f}{W_i} \right) \times 100 \]

\[ C_iR = \left( \frac{C_{i0} - C_{if}}{C_{i0}} \right) \times 100 \]

The percent of solid recovery (SR) is calculated from the dry weight of biomass, before (W_i) and after pretreatment (W_f). The percentage recovery of individual carbohydrates (C_R) for 3^th species (i.e., cellulose, hemicellulose or lignin) is calculated based on the amount present in dry biomass before (C_{i0}) and after (C_{if}) pretreatment. The compiled IL-based pretreatment dataset has been submitted to the Mendeley Data® repository with DOI: 10.17632/w3728zkf188.1.

2.1.2. Data imputation techniques

A significant proportion of predictor data (particularly on IL properties) was missing completely at random which can impart a significant degree of bias, reduce efficiency, and are not compatible with some standardized ML algorithm. Three different data imputation techniques, i.e. kNN imputation, regression tree (Jadhav et al., 2019), and principal component analysis based trimmed score regression (PCA-TSR) from missing data imputation (MDI) Toolbox (Folch-Fortuny et al., 2016) were considered.

For kNN imputation, a k-nearest neighbour classification model is built for the predictor to be imputed using fitcknn function in Matlab. A k-value of 11, i.e., the square root of the number of observations (123), was chosen as the number of nearest neighbours (Beretta & Santaniello, 2016). For regression tree-based imputation, a random forest ensemble of 200 regression trees is used to model the predictor with missing values. For PCA-TSR, the missing values are first filled with the column average of observed variables, mean-centred. Covariance and singular value decomposition on the matrix is performed, before reconstructing and constructing the matrix with imputed values. Further details on PCA-TSR can be found elsewhere (Folch-Fortuny et al., 2016).

All three methods were screened for their suitability for data imputation based on their capacity to “recover” ground-truth values of simulated missing data points (Waljee et al., 2013). For the same, 25% of known values from each predictor were randomly removed from the original data set. All missing values were imputed based on the imputation algorithm. The mean absolute percentage error (MAPE) of those artificially missing points for f^th predictor data set was calculated as shown in Eq. 3:

\[ MAPE_f = \frac{1}{m_f} \sum_{i=1}^{m_f} \left| \frac{x_{ij} - \hat{x}_{ij}}{x_{ij}} \right| \times 100 \]

Where, x_{ij}, \hat{x}_{ij} are the true and imputed values of f^th item of f^th predictor variable, respectively. The m_f is the total number of available data points treated as missing for f^th predictor. The random selection of data points to treat as “missing” was repeated 200 times, distribution of MAPE across the independent runs was used to compare the imputation efficiency of the algorithm.

2.2 Correlation and PCA of predictors and responses

The inclusion of predictors with linear/monotonic relationships (i.e., highly correlated features) can limit a model’s performance (Ascher et al., 2022). As the predictors relevant to IL or biomass properties are likely to be correlated, principal component analysis (PCA) method was used to project the data into a set of fewer uncorrelated (orthogonal) principal components (PC) while retaining much of the original variations. As the data set variables are of different units, the z-scores normalization (auto-scaling) was adopted before PCA analysis. The number of PCs to be retained was identified based on the % variance explained by them. PC with eigenvalues ≥ 1 retaining the significant variation of original data were only considered (Camargo, 2022). Varimax rotation of factors, i.e., maximizing loading on factors by individual parameters, was undertaken to increase their interpretation. PCA is performed using Statistics and Machine Learning Toolbox, ver. 12.5.

2.4 Artificial neural network (ANN) modelling
### 2.4.1. Optimization of HPs

Four independent fully connected feedforward neural network (ANN) architectures made up of an input layer, one/more hidden layer(s) and an output layer containing one of the four response variables \(i.e.,\) cellulose, hemicellulose, lignin recovery, and delignification was adopted for regression. The ANN model HPs, \(i.e.,\) number of hidden-layer and neurons, choice of activation function, L2 regularization strength, and standardization of predictors (true/false) were optimized using a Bayesian approach, which incorporates an acquisition function to determine the next point to evaluate and is much faster than grid search algorithm (Feurer & Hutter, 2019). However, a single Bayesian optimization run with wider HPs space boundary setting, \(i.e.,\) number of layers, and range of neurons in each layer, may not trace optimal network architecture. A series of HPO optimizations, each one with an increasing upper limit for the number of neurons (\(i.e.,\) 10, 30, 50, 70, and 90) in each layer for a maximum of two hidden layer configurations, were considered. The 10-fold cross-validation \(R^2\) from 25 different Bayesian optimization runs was used to map the model fitness landscape and trace the optimal model. The acquisition function was set to expected-improvement-plus, default exploration ratio of 0.5, with a maximum number of objective evaluations set to 60. The choice of cross-validation partitioning \(i.e.,\) seed for random number generator was kept unaltered in the exercise. The ANN model development was performed using the \texttt{fitnet} function from Statistics and Machine Learning Toolbox™ Ver. 12.5 (Matlab® R2023a, MathWorks).

### 2.4.2. Predictor selection model reduction

The importance of the predictors in the developed full-featured ANN model has been verified through permutation feature importance (Dudek, 2015) that measures change in mean square error (MSE) after randomly permuting one predictor while keeping all other features unchanged. A total of 100 different permutations on each of the predictors were repeated, and the average change in the MSE of the model was calculated for comparison.

The predictor selection combined with HPO was performed using a Bayesian optimizer, where all 13 predictors were also included as categorical optimizable variables. In each iteration, the selected predictors and HPs are passed on to the objective function to build a regression ANN. The cross-validation error of the developed model is set as an objective to minimize. In each iteration, the cross-validation partition is kept unaltered. However, as the optimization result \(i.e.,\) the variable selection and HPO can differ depending on the dataset composition of the cross-validation fold, the entire exercise is repeated 20 times. The frequency of the selected variables across the runs and their average fitness \(R^2_{CV}\) are reported. The variable selection-HPO scheme was adopted using a custom-made script in Matlab.

### 2.5. Random Forest Methods

RF is one of the most effective ensemble machine-learning regression models composed of a weighted combination of multiple regression trees (Sage et al., 2021), constructed for each of the tree, by random sampling with replacement (Bootstrapping) from sample data and random selection of predictors at node splitting step based on information gain (or Gini impurity) (Rodriguez-Galiano et al., 2014). At last, the above steps are repeated until the tree is complete. Predictions from all independent decision trees are averaged in an ensemble approach.

#### 2.5.1. HPs tuning for RF model and predictor of importance

The performance of the ensemble RF model depends on the associated HPs \(i.e.,\) the number of the decision tree in the forest \(ntree\), the maximal number of decision splits per tree \(nsplit\), minimum observations per leaf \(nleaf\), and the number of predictors to select at random for each split \(npred\). The first two HPs control the depth of regression trees. A small leaf size provides more splits, resulting in a deep tree, which may cause overfitting. In contrast, a large leaf size stops the tree from growing after one split or a few splits, resulting in poor predictive performance (Hossain et al., 2022).

Four independent RF models, for each of the responses in IL-based biomass pretreatment, were developed through the Bayesian HPO approach while minimizing five-fold cross-validation loss in objective evaluations. emerged as a practical tool for high-quality parameter selection in prediction systems. Bayesian HPO method internally maintains a Gaussian process model, and provides efficient alternatives to the grid or random search useful for optimizing black-box objective functions (Shahriari et al., 2016). The upper limit of optimizable variables was kept at their default, \(i.e.,\) 10–500 \(ntree\), 1–122 \(nsplit\), 1–61 \(nleaf\). As predictor importance analysis had been one of the aspects, all predictors were considered to be selected at random and \(npred\) was set to “all” in this HPO exercise.

Out-of-bag, predictor importance estimates by permutation measure how influential the predictor variables in the model are in predicting the response. For each regression tree, out-of-bag observations error is subtracted from model error using the out-of-bag observations containing the permuted values as a predictor (Genuer et al., 2010). The Average and standard deviation of the error over the all-regression tree for the specified predictor is computed. RF model development, training and optimization using the \texttt{fitrensemble} function from Statistics and Machine Learning Toolbox™ Ver. 12.5 (Matlab® R2023a, MathWorks).

#### 2.5.2. Predictor selection and HPO-optimized RF model

Predictor selection combined with HPO was performed by adopting three different approaches, forward selection, backward elimination (Xia & Yang, 2022), and integrated Bayesian optimization. Backward elimination starts with an HP-optimized RF model with full-predictor sets and importance measure ranking. Predictors are recursively eliminated based on their order of importance unless the \(R^2_{CV}\) of HP optimized model stops improving. Forward selection strategy starts with a null model and recursively adds up predictors in decreasing order of importance unless \(R^2_{CV}\) cease to improve (Nugroho et al., 2021). In the third strategy, \(i.e.,\) combined HPO-predictor section, all 13 variables with categorical encoding \(i.e.,\) true/false were also included as additional optimizable variables along with default HPs in Bayesian optimization.

### 2.6. Evaluation indicators
To quantitatively evaluate the performances of ANN and RF models, two common evaluation indicators, including the determination coefficients ($R^2$) and cross-validation $R^2$ ($R^2_{CV}$), are introduced and can be expressed as Eq. (4), and (5), respectively.

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]

\[ R^2_{CV} = 1 - \frac{\sum_{j=1}^{k} \sum_{i \in j} (y_{i,j} - M(y_{i,j}))^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]

Where, $y_i$, $\bar{y}$, $\hat{y}_i$, are experimental, experimental average, and model predicted values for $n$ number of observations. Cross-validation $y_{i,j}$ represents $i^{th}$ experimental point in the $j^{th}$ fold out of $k$-fold partitioning. $M(y_{i,j})$ is the prediction $y_{i,j}$ from the ML model (M) trained on $y_{i,j}$.

3. Results and discussion

3.1. The imputation method selection and statistics

The performance of the different data imputation techniques on simulated missing data is shown in Fig. 1. It is evident that across all three imputation methods, tree-based regression can predict the missing values with greater accuracy, i.e., with the lowest MAPE from 200 independent runs. The median MAPE of imputed data were very nominal (~10%) for predictors with a moderate degree of missingness, i.e. molecular weight, density and melting point with 11.38%-19.51% missing data. The narrow spread of MAPE suggests the reproducibility of the imputation methodology. Though the MAPE of imputed values for conductivity (with 46.34% of originally missing data) remains high, the performance of the RF imputation approach is better than the others. Studies have suggested RF imputation to be generally robust under moderate to high missingness, and even when data was missing not at random (Tang & Ishwaran, 2017). Similar results have also been reported in other studies, where the author observed the lowest MSE with RF out of nine candidate imputation methods in four different percentages of missing values of different origins (Kokla et al., 2019). A simple mean substitution approach has also been adopted in data imputation for the development of an ANN model predicting the biomass gasification process (Ascher et al., 2022). However, the impact of such imputation on the resulting distribution of data remains unknown. For predictors with a lower % of originally missing data, the choice of imputation appears to be less trivial as the range, mean, and standard deviation of imputed variables are very similar across the methods (Table 1). However, average conductivity and viscosity (originally 46.34% and 37.40% missing data, respectively) after imputation differ across the methods.
3.2. Predictor space correlation and PCA analysis

The predictors considered for the IL-based pretreatment of biomass in this study are of three different domains, i.e., IL properties, biomass characteristics and pretreatment conditions. The Pearson correlation across the predictor space is shown in Table 2. The existence of any natural association across the entire predictor set is very unlikely, however, the IL properties may be correlated with one another. For example, the density of IL shows a strong and significant association with conductivity ($p < 0.05$, $r = -0.765$) and melting points ($p < 0.05$, $r = 0.787$). For biomass characteristics, cellulose and hemicellulose content are strongly correlated with other extractives ($p < 0.05$, $r = -0.787$) and lignin content ($p < 0.05$, $r = 0.573$), respectively. This seems very logical as the proportion of all those biomass components adds up to unity and is linearly dependent (Jolliffe & Cadima, 2016). As expected, there is no significant relationship between the user-designed treatment conditions.

The extent of variability across the dataset is reflected in the z-score boxplot (Fig. 2a). The box indicates the 25th and 75th percentiles of the values, whereas the whisker extends to data points, still not considered outliers. Most of the predictors are normally distributed, though few outliers are evident in all cases. As only a small subset of predictors is highly correlated, an increasing number of PCs (not just one/two) would be required to explain a reasonable amount of information. Figure 2 (b) demonstrates a scree plot of every component's contribution toward the total variance of the dataset, revealing 5 PCs with eigenvalues greater than one, accounting for 76.3% of the entire variance in the dataset, and 9 PCs are required to capture 95% of the total variance. Figure 2 (c) shows the relationships between the first two PCs that contribute the most to the overall variance. The loading of the parameters is represented by a vector,
while the scores of the samples are shown in a scatter plot. Only their relative locations can be established because the scores are scaled about the highest score value and highest coefficient length. The direction and magnitude of the vector reveal how much each variable contributes to the two principal components. For example, the first principal component, on the horizontal axis, has positive coefficients for density and melting point of IL and is negatively related to other extractives in biomass, and reaction time of pretreatment. On the vertical axis, the second major component exhibits positive coefficients for the hemicellulose content of biomass.

Table 2. Correlation matrix of data on IL pretreatment of biomass, involving IL properties, biomass characteristics and pretreatment condition. Significant correlations (p < 0.05) and strong associations (r > 0.5) are shown in boldface, and highlighted, respectively.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.194</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.205</td>
<td>-0.344</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.020</td>
<td>-0.767</td>
<td>0.332</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-0.223</td>
<td>0.750</td>
<td>-0.444</td>
<td>-0.751</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-0.160</td>
<td>0.127</td>
<td>-0.121</td>
<td>-0.048</td>
<td>0.046</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-0.137</td>
<td>-0.342</td>
<td>0.182</td>
<td>0.344</td>
<td>-0.349</td>
<td>0.047</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-0.359</td>
<td>-0.058</td>
<td>0.168</td>
<td>-0.112</td>
<td>0.168</td>
<td>-0.136</td>
<td>0.198</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.300</td>
<td>-0.075</td>
<td>0.152</td>
<td>0.113</td>
<td>-0.186</td>
<td>-0.074</td>
<td>-0.014</td>
<td>-0.573</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>0.122</td>
<td>0.327</td>
<td>-0.338</td>
<td>-0.240</td>
<td>0.256</td>
<td>0.107</td>
<td>-0.787</td>
<td>-0.376</td>
<td>-0.324</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-0.027</td>
<td>-0.296</td>
<td>-0.017</td>
<td>0.361</td>
<td>-0.393</td>
<td>0.067</td>
<td>0.417</td>
<td>-0.116</td>
<td>0.147</td>
<td>-0.310</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>0.507</td>
<td>0.358</td>
<td>-0.148</td>
<td>-0.162</td>
<td>0.096</td>
<td>0.032</td>
<td>-0.174</td>
<td>-0.209</td>
<td>0.035</td>
<td>0.237</td>
<td>-0.213</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>-0.320</td>
<td>0.252</td>
<td>0.107</td>
<td>0.137</td>
<td>-0.008</td>
<td>-0.111</td>
<td>0.339</td>
<td>0.357</td>
<td>-0.120</td>
<td>-0.379</td>
<td>0.063</td>
<td>-0.261</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*a* A: molecular weight, B: density (g/cm³), C: viscosity, D: conductivity, E: melting point, F: particle size, G: cellulose content, H: hemicellulose content, I: lignin content, J: other extractive in biomass; K: reaction temperature (K), L: reaction time (min), M: solid to liquid ration (w/w). A statistically significant correlation (p < 0.05) between predictors is shown in boldface.

### 3.3. ANN-based modelling of IL pretreatment

#### 3.3.1. Optimization of HPs

The outcome of 25 independent Bayesian optimization runs, with different combinations of hidden layer neuron's upper limits, for each of the four independent responses (i.e., solid, cellulose, hemicellulose, and lignin recovery) prediction is shown in Fig. 3. The fitness of ANN models projected from Bayesian optimization varies appreciably depending on the imposed upper limits of the number of neurons considered in bilayer architecture.

Increasing the upper limit of neurons in HPO optimization runs can not guarantee the identification of best-fit models. Except for solid recovery, the region of high fitness is scattered across the entire landscape. The best cross-validation R² for cellulose recovery is superior to the two other responses. In ANN modelling of fuel properties for waste solid biomass hydrochar, it has been observed that model fitness (training R²) generally increases with the number of neurons in single-hidden layer architecture (Mu et al., 2022). However, such an assessment could be criticized for maintaining training R² as an evaluation criterion. Researchers have also adopted a grid search approach as straightforward and exhaustive to explore through all the combinations of manually specified HPs in ANN, or RF models (Huang et al., 2023). Though the exploration of only three HPs at just three levels in the referred study would not have been overwhelming, Bayesian optimisation is certainly a better choice for efficient global optimisation of expensive architecture as in this case (Greenhill et al., 2020).

The best fit ANN models from 25 Bayesian HPO runs for each response are shown in Table 3. All the available activation functions, i.e., sigmoid, tanh, ReLu are selected across the responses. The large number of neurons retained in the 1st and 2nd layers can be attributed to the larger predictor dimension. Unlike the two-step HPOs optimization approaches, exhaustive grid search exploring different combinations of the number of neurons in 1st and 2nd hidden layers, data split for the test set, training function, and hidden layer transfer function has been adopted in literature (Ascher et al., 2022). The fitness of the model remains exceptionally good as the regression coefficient (R²) training dataset varies between 0.936 and 0.993 (Fig. 4) depending on the responses. However, cross-validation R² for cellulose recovery (0.571) is not very promising, possibly implicating poor generalization capability of the ANN models. Though ANN models developed with a large number of predictors are prone to overfitting, the use of R²adj statistics to penalize excessive predictors as adopted in many studies (Ascher et al., 2022), may not be equally powerful as to cross-validation score. It should be noted that a reasonably large data set with homogeneous unbiased coverage of predictors is important to achieve good generalizability (Zhang et al., 2023). However, in the compilation of data from independent research, such a criterion could not be enforced.
Table 3
HPO result for the regression ANN predicting different responses from IL-based biomass pretreatment

<table>
<thead>
<tr>
<th>Response</th>
<th>Activation</th>
<th>Lambda</th>
<th>Number of neurons</th>
<th>Model fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Layer_1</td>
<td>Layer_2</td>
</tr>
<tr>
<td>Solid recovery</td>
<td>sigmoid</td>
<td>0.0030</td>
<td>49</td>
<td>-</td>
</tr>
<tr>
<td>Cellulose recovery</td>
<td>tanh</td>
<td>0.0214</td>
<td>82</td>
<td>-</td>
</tr>
<tr>
<td>Hemicellulose recovery</td>
<td>sigmoid</td>
<td>0.0226</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Lignin recovery</td>
<td>reLu</td>
<td>0.3610</td>
<td>39</td>
<td>3</td>
</tr>
</tbody>
</table>

3.3.2. Predictor selection and dimension reduction

The permuted predictor’s importance in the full-feature ANN models is shown in Fig. 5. Except for the ANN model predicting hemicellulose content, temperature and time of IL treatment remain one of the most important predictors. Melting point and density are the most noteworthy predictors from the IL properties subgroup for the ANN model predicting cellulose recovery. Studies have suggested that the density of ILs impacts cellulose-dissolving capability (Abe et al., 2015). Though the importance of IL conductivity in the electro-assisted pretreatment of lignocellulosic materials is well understood (Sun et al., 2020), the significance of conductivity in thermal pretreatment is not evident in the literature. Cellulose and hemicellulose content are the most influential predictors from the biomass characteristics group for the cellulose and hemicellulose recovery model. Though the relative importance of predictors can be easily interpreted, this model-agnostic method approach does not disregard the feature altogether from the model (Molnar, 2020), and the fitness of the reduced dimension is not evaluated.

The variable selection integrated with HPO across 20 independent repetitions is shown in Fig. 6. Though no feature selection algorithm can guarantee a globally optimal solution, temperature and time are the most frequently selected variables in ANN models predicting solid, cellulose recovery. In addition, lignin content appears to be very crucial in the lignin recovery model (Fig. 6d). The importance of solid-to-liquid ratio in hemicellulose recovery (Fig. 6c) is well documented (Gallardo et al., 2020). The propensity of the selected feature is indeed very similar to the findings in permuted feature importance analysis. The cross-validation R^2 of the ANN models with reduced parametric space are certainly better than their full-predictor counterparts.

3.4. Random forest Modeling

3.4.1. Full predictor-optimized model

The outcome of HPO for each of the bagged random forest models, retaining all predictors, and targeting different responses is shown in Table 4. It is evident that, except for hemicellulose recovery, the cross-validation performance of the developed RF models is significantly lower. Even though the upper limit for the number of regression trees was set at 500 in Bayesian optimization, the optimally retained number of trees in RF is indeed very few. Though a lower number of trees can have limited prediction power, increasing the number of trees does not necessarily improve model quality further (Torre-Tojal et al., 2022). On the other hand, a high value of ‘maximum number of splits’ captures more intricate patterns in the data, potentially leading to a more expressive model. It should be noted that in RF regression considering “all” variables to decide each split, the contribution from less important variables, which otherwise could be useful for the prediction of a small group of observations, gets “masked” by stronger predictors (Probst et al., 2019).

Table 4
The optimized HPs for random forest regression models (with full predictor sets) used for IL-based pretreatment of biomass

<table>
<thead>
<tr>
<th>Response</th>
<th>HPs a</th>
<th>Fitness b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ntree</td>
<td>nleaf</td>
</tr>
<tr>
<td>Solid recovery</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>Cellulose recovery</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td>Hemicellulose recovery</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>Lignin recovery</td>
<td>79</td>
<td>2</td>
</tr>
</tbody>
</table>
The fit of the RF model predictions with experimental data is shown in Fig. 7. The regression coefficient for the ensemble model predicting solid and hemicellulose recovery remains reasonably good ($R^2 = 0.824–0.939$). There is no specific pattern or trend observed in deviations for the models, or the errors are normally distributed.

The normalized out-of-bag predictor importance for the RF models, considering “all” variables to make a decision split, is shown in Fig. 8. The OBB importance of a few predictors, e.g., reaction temperature, hemicellulose and lignin content for solid recovery, hemicellulose, and lignin recovery, respectively are distinctively high in comparison to remaining subsets in RF models. The most important predictors identified are certainly reasonable for respective responses. It is obvious that all input variables, with different importance, are contributing to the prediction of all the responses. Similar observations have also been made in other studies (Zhang et al., 2019), where no variables could be dropped from the final RF model. However, the decision to retain all variables, have a model reduction or even the choice of model reduction strategy (i.e., forward, backward, or population-based) can be subjective.

### 3.4.4. RF model with reduced predictors

The predictor selection frequency across 20 independent runs of PS-HPO, using a Bayesian optimization scheme, is shown in Fig. 9. Molecular weight, melting point of IL, particle size, lignin, and other extractive contents in biomass are mostly excluded in the solid recovery model. The melting point of IL and particle size are rarely selected in the variable selection exercise across all four RF models for different responses. The $R^2_{CV}$ across the twenty independent RF models varied between 0.543–0.681, 0.303–0.474, 0.786–0.824, and 0.541–0.582 for solid, cellulose, hemicellulose, and lignin recovery. The fitness of RF models with optimzed predictor subsets is superior to their full-predictor RF model variants.

The outcome of forward, and backward variable selection (using the OBB predictor importance) in comparison to the Bayesian optimized variable selection approach in RF regression is shown in Table 5. It is clear that OBB predictor importance-guided backward elimination is not effective in reducing the variable space. On the other hand, the forward selection strategy incorporates a very limited subset of predictors in RF models (e.g., hemicellulose and lignin recovery), impacting their predictive and generalization capability. The size of the selected predictor set in Bayesian optimization, however, lies between the two extremities with consistently better $R^2_{CV}$ than the other two approaches.

### 3.5. Comparison of the ML modelling approaches for the data sets

The performance of the ANN model predicting IL-pretreatment efficiency remains significantly better than their RF counterparts on training observation ($R^2$). The cross-validation performance of the ANN models (0.657, 0.547, 0.761, 0.658) was also superior to that of RF models (0.554, 0.383, 0.831, 0.513) considering the full-predictor set. Following the Bayesian predictor selection, the ANN model fitness (best $R^2_{CV}$) predicting solid, cellulose, hemicellulose, and lignin recovery improved to 0.736, 0.555, 0.825, and 0.651, respectively. Predictor selection in the RF model on the other hand improved the $R^2_{CV}$ to 0.681, 0.475, 0.824, 0.582, respectively. This assertion reveals an equivalent or marginally better performance of ANN over RF.

<table>
<thead>
<tr>
<th>Response</th>
<th>Method a</th>
<th>Retained variables ID</th>
<th>HPs b</th>
<th>Fitness b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ntree</td>
<td>nleaf</td>
<td>nsplit</td>
<td>nvar</td>
</tr>
<tr>
<td>Solid recovery</td>
<td>BS</td>
<td>{1,2,3,5,6,7,9,10,11,12,13}</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>Cellulose recovery</td>
<td>BS</td>
<td>{1,2,3,4,5,6,7,8,9,10,11,12}</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>Hemicellulose recovery</td>
<td>BS</td>
<td>{1,2,3,4,6,7,8,9,10,11,12}</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>Lignin recovery</td>
<td>BS</td>
<td>{2,4,5,6,7,8,9,10,11,12,13}</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Solid recovery</td>
<td>FS</td>
<td>{3,11,12}</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Cellulose recovery</td>
<td>FS</td>
<td>{7,9,11}</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Hemicellulose recovery</td>
<td>FS</td>
<td>{8}</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Lignin recovery</td>
<td>FS</td>
<td>{9}</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Solid recovery</td>
<td>BO</td>
<td>{4,5,7,8,11,12}</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>Cellulose recovery</td>
<td>BO</td>
<td>{1,2,3,4,5,7,8,9,10,11,12}</td>
<td>118</td>
<td>3</td>
</tr>
<tr>
<td>Hemicellulose recovery</td>
<td>BO</td>
<td>{3,4,6,7,8,9,10,11,12,13}</td>
<td>228</td>
<td>2</td>
</tr>
<tr>
<td>Lignin recovery</td>
<td>BO</td>
<td>{1,2,3,6,8,9,11,12}</td>
<td>114</td>
<td>1</td>
</tr>
</tbody>
</table>

a ntree – number of tree in RF, nleaf – minimum nos. of observations per leaf, nsplit – maximum nos. of split. b RMSE: root means square error of the regression ensemble, $R^2$ and $R^2_{CV}$ – regression coefficient of the trained and 5-fold cross-validated models, respectively.
Method: BS- backward elimination, FS- forward selection based on the OBB predictor of importance data. BO- Bayesian optimization results shown here are the best outcome (in terms of $R^2_{CV}$) out of twenty repetitions. \( \text{*ntree - number of learning cycles, nleaf - minimum number of observations per leaf in regression tree, nsplit - maximum number of split, nvar - number of variables randomly selected at each node. Selected predictors 1–13 refer to (1) molecular weight, (2) Density, (3) viscosity, (4) conductivity, (5) melting point of IL, (6) particle size, (7) cellulose, (8) hemicellulose, (9) lignin, (10) other extractive contents in biomass; (11) temperature, (12) Time, (13) solid to liquid ratio in pretreatment step.}

4. Conclusion

The influence of biomass characteristics, properties of ILs, and processing conditions on solid, cellulose, hemicellulose, and lignin recovery in IL-based LCB pretreatment was studied using ANN and RF-based modelling approaches. Compilation of heterogeneous experimental data from literature and their pre-processing (including missing value imputation technique) is certainly an important issue. Among the others, temperature appeared to be the most important variable affecting the process of pretreatment. The density and conductivity of ILs were also important in biomass pretreatment. The prediction performance of hyperparameter-optimized ANN remains superior to the RF variant. Bayesian predictor selection combined with HPO offered improvement in generalization capability. Collection and curation of unbiased homogeneous data is warranted to improve the performance of ML adopted for biomass pretreatment.

Declarations

Acknowledgements

DH and BM would like to acknowledge Karunya Institute of Technology and Sciences, Coimbatore for providing every essential support to perform the experiments and complete this research work.

Ethics approval and consent to participate

The study does not involve any human participants, human data or human tissue. Ethics approval is not applicable.

Consent for publication

Not Applicable.

Availability of data and materials

Data in this study are available on request to the corresponding author.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

The authors have no relevant financial or non-financial interests to disclose.

Funding

This work is financially supported by Karunya Institute of Technology and Sciences, Coimbatore.

Authors contributions statement


References


Figures

Figure 1

The distribution of MAPE across 200 independent runs for each predictor using different imputation methods. The assessment is based on imputed values of randomly selected 25% data points for each predictor when simulated as missing. Predictors: $P_1$-molecular weight, $P_2$-density, $P_3$-viscosity, $P_4$-conductivity, and $P_5$-melting point. The reference line of 10% and 20% MAPE are shown for ease comparison.
Figure 2

(a) Z-score variability of 13 independent predictors for IL pretreatment data set. (b) The scree plot shows the % variance explained by PCs. The first 5 PCs with eigenvalues >1 account for 76.30% variance, while 9 PCs are required to have 95% variance. (c) Biplot showing sample scores and variable loading onto the first two principal components. Predictors: A: molecular weight, B: density (g/cm$^3$), C: viscosity, D: conductivity, E: melting point, F: particle size, G: cellulose content, H: hemicellulose content, I: lignin content, J: other extractive in biomass; K: reaction temperature (K), L: reaction time (min), M: solid to liquid ratio (w/w).
Figure 3

Influence of upper limit setting for the number of neurones in Bayesian optimization of bilayer ANN modelling architecture for (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery in IL pretreatment dataset. Model fitness (i.e., cross-validation $R^2$) heat map is shown in color bar.
Figure 4

Regression scatter plot between experimental and ANN modelled response for (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery.
Figure 5

Permuted predictors of importance in ANN models for (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery. An increase in model MSE (average of 100 independent repetitions), in comparison to the reference model, due to permuted values of a predictor signifies the importance. Predictor abbreviation: Mol – molecular weight, Den – Density, Vis – viscosity, Con – conductivity, MP – melting point of IL; Size – particle size, Cel – cellulose, Hem – hemicellulose, Lig – lignin, Ext – other extractive contents in biomass; Temp – temperature, Time – Time, SLR – solid to liquid ration in pretreatment step.
Heatmap of selected (dark cells) and dropped (light cells) variables across 20 independent Bayesian predictor selection exercises in ANN models for (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery. Predictor abbreviation: v1 – molecular weight, v2 – Density, v3 – viscosity, v4 – conductivity, v5 – melting point of IL; v6 – particle size, v7 – cellulose, v8 – hemicellulose, v9 – lignin, v10 – other extractive contents in biomass; v11 – temperature, v12 – Time, v13 – solid to liquid ration in pretreatment step.
Figure 7

Regression scatter plot between experimental and RF modelled response (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery.
Figure 8
Heatmap showing normalized OBB predictor importance of the RF models.

Figure 9
Heatmap of selected (dark cells) and dropped (light cells) variables across 20 independent predictor selection exercises in RF models for (a) solid, (b) cellulose, (c) hemicellulose, and (d) lignin recovery. Predictor abbreviation: v1 – molecular weight, v2 – Density, v3 – viscosity, v4 – conductivity, v5 – melting point of IL; v6 – particle size, v7 – cellulose, v8 – hemicellulose, v9 – lignin, v10 – other extractive contents in biomass; v11 – temperature, v12 – Time, v13 – solid to liquid ration in pretreatment step.