Bioresource Technology Reports 18 (2022) 101027

Contents lists available at ScienceDirect

Bioresource Technology Reports

journal homepage: www.sciencedirect.com/journal/bioresource-technology-reports

Machine learning prediction of SCOBY cellulose yield from Kombucha tea fermentation

Thangaraj Priyadharshini [a], Krishnamoorthy Nageshwari [a], Senthamizhan Vimaladhasan [b], Sutar Parag Prakash [c], Paramasivan Balasubramanian [a,*]

[a] Department of Biotechnology & Medical Engineering, National Institute of Technology Rourkela, 769008, India
[b] Centre for Biotechnology, Anna University, 600025, India
[c] Department of Food Process Engineering, National Institute of Technology Rourkela, 769008, India

ARTICLE INFO

Keywords:
Kombucha biofilm
Bacterial cellulose yield
Machine learning
Extreme gradient boosting
Predictive modeling

ABSTRACT

The commercialization of by-product of Kombucha SCOBY (Symbiotic consortium of bacteria and yeast) could be the sustainable way of transforming waste into value-added products. This study aims at developing a robust machine learning model for the prediction of SCOBY yield. Concentrations of tea, sucrose, SCOBY, inoculum, pH, temperature, and fermentation time were the input parameters considered. The robustness of the models was studied using correlation coefficient and root mean square error. Among the algorithms studied, XGB (eXtreme Gradient Boosting) was the most resilient model with high accuracy. By hyperparameter tuning and *k*-fold cross-validations, the model performance was improved to attain an R^2 value of 0.9048. The relationship between variables was depicted as Pair plot and Pearson correlation matrix. Fermentation temperature was the most influential parameter affecting SCOBY yield. Shapley additive explanations dependence plots and summary plot provided insights on the combined effects of input parameters on the SCOBY yield.

1. Introduction

In nature, cellulose is the most occurring polymer obtained from various sources namely, plants, bacteria, algae, etc. The unique physicochemical properties and biodegradable nature of cellulose have attained the limelight among researchers, and thus, the biopolymer is continuously being exploited to bring out the prospective findings (de Oliveira Barud et al., 2016). Cellulose is the main component of the plant cell wall and is found along with hemicellulose, lignin, and pectin. Though plant-based cellulose is available in abundance, more focus is bestowed on bacterial cellulose that has superior features as high chemical purity, high water holding capacity, high crystallinity, and hydrophilicity (Naomi et al., 2020). Besides, bacterial cellulose extraction is an eco-friendly and convenient process than other conventional methods used for extraction from woods or cotton that requires more energy and harmful chemicals (Kamiński et al., 2020).

Acetic acid bacteria synthesize cellulose as an extracellular polysaccharide by microbial fermentation (Chakravorty et al., 2019). Symbiotic growth of certain bacteria and yeast species is found to be the most efficient system for cellulose production. Kombucha fermentation is one such popular process practiced, involving the symbiotic consortium of bacteria and yeast (SCOBY) (Kamiński et al., 2020; Markov et al., 2001; Sharma and Bhardwaj, 2019). The fermentation process is initiated by the inoculation of live mother SCOBY to the sugared tea broth. Following that, kombucha tea is added and incubated at room temperature for few weeks (Jayabalan et al., 2007). Kombucha tea or beverage produced during the fermentation is the primary product of choice, with SCOBY biofilm rich in cellulose generated as the by-product (Gargey et al., 2019). Bacterial species present dominantly in the SCOBY are members of the acetic acid bacteria family (*Komagataeibacter xylinus*, *K. intermedius*, *Acetobacter aceti*, *A. pasteurianus*, and *Gluconobacter oxydans*) and lactic acid bacteria (*Lactobacillus* and *Lactococcus*). *Saccharomyces*, *Zygosaccharomyces*, *Brettanomyces*, and *Pichia* are the common yeast species found in the SCOBY (Chakravorty et al., 2019; Mukadam et al., 2016; Treviño-Garza et al., 2020). SCOBY containing microbial community adhering to it can be an inoculant for the subsequent batches (Jayabalan et al., 2010). Other than that, the SCOBY biofilm remains unutilized in most cases; however, materializing SCOBY would be an ideal means of high returns with less capital investment (Sharma and Bhardwaj, 2019). Recently, SCOBY has found application

* Corresponding author.
E-mail address: biobala@nitrkl.ac.in (P. Balasubramanian).

https://doi.org/10.1016/j.biteb.2022.101027
Received 7 February 2022; Received in revised form 15 March 2022; Accepted 16 March 2022
Available online 28 March 2022

in several fields such as food packaging industry, tissue engineering, bioleaching agent, animal feed, textile industry, electronic batteries, treatment of cancer and as a biosorption material (Laavanya et al., 2021). The increasing demand for bacterial cellulose and the idea of additional profits have kindled the concept of proper consumption of kombucha SCOBY (Treviño-Garza et al., 2020).

Optimization of fermentation will aid in attaining better SCOBY yield with limited substrates. So far, the conventional statistical modeling employed for optimization involve performing trial runs under the range of production conditions that is time-consuming and requires more resources. To achieve maximum SCOBY yield, few experimental studies have been conducted to determine the key process parameters in kombucha fermentation. Various factors reported to influence the SCOBY yield are the choice and quantity of substrate, inoculation method, incubation temperature, duration of fermentation, initial pH, type of SCOBY inoculum, and the dimensions of culture vessel (Abd El-Salam, 2012; Soh and Lee, 2002; Treviño-Garza et al., 2020). For example, the type of tea extract and its concentration used tends to affect the yield and thickness of the cellulose. The operational parameters such as pH and fermentation temperature primarily favor selective growth of distinct bacterial species (De Filippis et al., 2018). The above varying conditions ultimately affect the microbial community of SCOBY by means of growth rate or metabolic activity. Thus, the synthesis of bacterial cellulose and other metabolites from SCOBY is an intricate process (Laavanya et al., 2021). Given the complexity of the process, it demands the development of models that associate key input factors to the SCOBY biofilm yield. The implementation of machine learning possibly will unravel the underlying complex patterns and trends among input and output parameters.

Machine learning is a futuristic modeling technique used for the optimization and scaling-up processes. It utilizes data from available literature, offering a time and energy-saving approach (Weichert et al., 2019). In traditional optimization processes, the objective of the model is to find the global maxima to increase the production. In addition, the maxima identification is done only for the data at hand (Balkanski et al., 2017). As opposed to that, machine learning tries to identify a generalized solution to the problem to ensure that the model remains robust to future perturbations (Jordan and Mitchell, 2015). It uses stochastic optimization to ensure robustness and this characteristic is particularly important in biological production systems, which have a lot of complex reactions and unknown factors that can influence the end results. A supervised machine learning model works by developing an algorithm based on the input training data to predict the desired output, where the robustness of the model depends on the size of the datasets (Volk et al., 2020). Machine learning takes fewer assumptions compared to conventional statistical methodologies such as Design of Experiments (DOE) using Analysis of Variance (ANOVA), which assumes that the residuals are independent, random and normally distributed. The amount of data required to construct a robust experimental design is far greater compared to what machine learning requires. Also, the critical features needed to successfully employ DOE are not always easy to evaluate and obtain. Machine learning overcomes such limitations and enables more accurate predictions using fewer features and samples (Laberge, 2011). In this study, it shown that machine learning methods can be used to build a better prediction model without making any prior assumptions.

The application of predictive modeling for industrial fermentation can improve resource utilization and provide an understanding of how the microbiota functions and progress in different conditions (Wang et al., 2021). Furthermore, the fermentation performance analysis could be performed by predictive modeling to maximize the yield. A diverse class of supervised machine learning algorithms has been developed and can be applied to deal with various problems. The factors to be considered for selection of appropriate model are, the problem to be addressed, size and quality of data, accuracy, desired output and, interpretation requirements (Goodswen et al., 2021). Therefore, the promising machine learning models selected for the prediction study of SCOBY yield are: Linear regression model, Polynomial regression model, and tree based models such as ExtraTrees, Random Forest and XGB (eXtreme Gradient Boosting). Regression models are the classic tools used in applications of prediction by determining the relationship between the input and output parameters (Maulud and Abdulazeez, 2020). Decision tree models can be exploited for both classification and regression problems. The algorithm works by splitting of the training data continuously based on information gain that selects one variable each on every node split (Lee et al., 2020). Hence, the additional feature of decision trees is that the significant variable from which the key decisions are made could be obtained. Goodswen et al. (2021) describes the standard steps for developing a supervised machine learning model in detail. The objective of this study is to utilize machine learning model for comprehending the relationship between the input and output variables for the prediction of model output.

To the best of our knowledge, no previous studies have been conducted on developing a machine learning model for the kombucha SCOBY yield. This manuscript aims to develop a robust machine learning model for the prediction of bacterial cellulose yield from kombucha. The input features are concentrations of tea, sugar, SCOBY, and kombucha inoculum, initial pH, temperature, and duration of fermentation. The individual and the combined effects of the input features on the outcome is elucidated from the SHAP (SHapley Additive Explanations) dependence plot and summary plot analysis and explained in detail. Further, the impact of the commonly used tea variety such as black tea, green tea, waste tea, and the mixture of green and black tea (2:1) on the yield has also been studied.

2. Materials and methods

2.1. Data collection and pre-processing

The first and foremost step in machine learning is the development of a well-structured database. A total of 323 datasets were collected from the tables, graphs, and Supplementary materials of the 30 existing literature published pertaining to kombucha fermentation. The input parameters under consideration were the process variables such as mass of tea leaves, concentration of sugar, mass of SCOBY inoculum, and volume of kombucha tea. Further, operating conditions of fermentation process such as pH, duration and temperature of the fermentation medium were also considered. Four types of tea extracts (Black tea, Green tea, Waste tea, and Black + Green tea) were included as categorical variables in the dataset by performing *one-hot encoding*. The targeted output variables were SCOBY yield and the overall productivity. However, it is noteworthy to mention that the SCOBY yield and productivity were provided considered on the wet weight basis for this study as most of the published literature data existed on wet basis. Yet, in few cases, where only the dry weight of the cellulose yield was reported, it the same was converted to wet weight basis with references to the previous studies (Muhialdin et al., 2019; Sharma and Bhardwaj, 2019).

Initial exploratory data analysis (EDA) was performed by imputing the missing data (unavailability of information) using the arithmetic mean, followed by data standardization using the *sklearn* package. Other EDA parameters included were the count, mean, standard deviation, and quartile values for all the input and output parameters. Data preprocessing is necessary for the successful development of a machine learning. In this process, few extreme datasets (outliers) affecting the model outcome and reliability were removed from the study.

2.2. Machine learning model selection

Several machine learning models are available for the users, each having its specific features and importance. Depending on the type of data available and model performance, the selection is finalized. Classification and regression are the two functions of supervised learning, where the former refers to discrete outcomes (qualitative) and the latter

Table 1
Statistical analysis of input and output model parameters for SCOBY yield from kombucha tea fermentation.

	Tea (g/L)	Sugar (g/L)	SCOBY (g/L)	Kombucha tea (ml/L)	pH	Temperature (°C)	Duration (days)	Yield (g/L)	Productivity (g/L·day)
Count	317	317	309	276	227	296	323	323	323
Mean	18.21	105.77	17.92	88.77	3.80	27.38	13.50	55.05	4.79
Standard deviation	33.88	101.41	11.73	63.78	0.95	3.92	8.64	75.05	6.41
Minimum	0.00	0.00	0.00	0.00	2.00	20.00	1.00	0.38	0.02
25%	5.00	50.00	8.00	50.00	3.00	25.00	7.00	13.19	1.12
50%	8.10	80.00	20.00	100.00	3.60	28.00	14.00	25.00	2.61
75%	12.00	100.00	30.00	100.00	4.00	30.00	15.00	64.79	5.19
Maximum	200.00	550.00	40.00	250.00	7.60	45.00	56.00	456.00	41.31

to the continuous actual outcomes (quantitative) (De Clercq et al., 2020). In this study, regression models were preferred for the prediction of SCOBY yield. Supervised machine learning algorithms employed in the study were: Linear regression model, Polynomial regression model, ExtraTrees regression model, Random Forest model, and eXtreme Gradient Boosting (XGB) model. The machine learning model selection and validation were executed using the *sklearn* 0.24.1 version. The datasets were standardized to range from 0 to 1 using *MinMaxScaler* tool in the software.

2.2.1. Linear regression

Linear regression works under the assumption of a linear relationship between input and output variables (Maulud and Abdulazeez, 2020; Volk et al., 2020). Here, multiple linear regression was preferred as the number of input features were more than one. The loss function, ordinary least squares, was used to minimize the sum of squared residuals.

2.2.2. Polynomial regression

The *PolynomialFeatures* function in *sklearn* package was used to assess the non-linear relationship between the input and output variables. Assigning of higher degree may also cause overfitting. Hence, a 2-degree polynomial was fixed considering the complexity of the datasets and the input-output relationship. The same loss function was used in both linear and polynomial regression models (Ordinary least squares). Loss function helps in determining the difference between actual and predicted values (Maulud and Abdulazeez, 2020).

2.2.3. Tree-based regression

Three tree-based models, ExtraTrees, Random Forest and XGB, all based on ensemble technique of decision tree combinations, were tested. The techniques involved in these models were of two classes based on how the decision trees are made: bagging (Parallel method) and boosting (Sequential method). In bagging method, many base functions are formed from the bootstrap samples and combined together to predict the output. ExtraTrees and Random Forest models use bagging, also known as bootstrapping. ExtraTrees known as Extremely randomized trees are derived from Random Forest. The difference between them is that the former chooses random cut-points from the entire sample for decision tree splitting, while in the latter, the cut-points are selected from a bootstrap replica (subsets) (Lee et al., 2020). XGB is a boosting ensemble algorithm that develops week classifier for each iteration, and subsequently, those are combined into a strong learner with more accurate predictions (Wang et al., 2019). The attractive feature of XGB such as fast programming speed, excellent model performance, high accuracy and efficiency are responsible for employing them extensively in many applications. The object function of XGB is to form K regression tree (K denotes number of trees) so that high accuracy is possible and high generalization capability (Wang et al., 2019). Though XGB is powerful and precise, there is a high possibility of overfitting (Pathy et al., 2020). This concern may be rectified by optimizing the hyperparameter tunings such as *learning_rate*, *max_depth*, *sub_sample*, *reg_lambda* and *reg_alpha*.

2.3. Model evaluation

Evaluation of the models is crucial in understanding the robustness of the model. The metrics used to select efficient regression models were correlation coefficient (R^2) and root mean square error (*RMSE*) values. R^2 indicates the goodness of fit of the developed models. *RMSE* is a measure of the standard deviation of prediction errors. Higher R^2 and lower *RMSE* values will help in identifying the best model. Thus, these metrics were calculated according to the following equations Eqs. (1) and (2) and compared for all five model candidates (Liu et al., 2019; Pathy et al., 2020),

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} \left(Y_i^{exp} - Y_i^{pred} \right)^2}{\sum_{i=1}^{N} \left(Y_i^{exp} - Y_{ave}^{-exp} \right)^2} \right) \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i^{pred} - Y_i^{exp} \right)^2}{N}} \quad (2)$$

where, N is the number of data points, Y_i^{exp} is the experimental values, Y_i^{pred} is the model predicted values, and $Y_{ave}{}^{-exp}$ is the average of the experimental values.

Further, the selected model performance was evaluated and studied using learning curve and actual (experimental) vs. predicted value scatter plot.

2.4. Feature importance and SHAP plots

Feature importance, dependence, and summary plots are some of the global model interpretation methods. Feature importance is crucial for understanding the relative significance among the input parameters and with the model output. If a feature is used relatively higher than others in making critical decisions, it will be considered to have a high significance on the output. The feature importance can be calculated based on different factors, such as weight, gain, and cover, where each aspect can provide varying results. Thus, the user has to choose according to the data and the required outcome of the study. In this study, feature importance was measured using the metric, total gain. SHAP dependence plots and the summary plot were plotted using *TreeExplainer*. SHAP values help in the thorough interpretation of the input-output variable relationship.

3. Results and discussion

3.1. Statistical and exploratory data analysis (EDA)

The final database consisted of 323 datasets from 30 literary articles. The seven input variables presented in the study were concentrations of tea, sugar, SCOBY, and kombucha tea inoculum, initial pH, temperature, and fermentation time. SCOBY yield and productivity were the desired

output variables. However, due to trend similarities and direct correlation, only biofilm yield was considered as output for analysis. The results obtained can be extrapolated for productivity as well. The input parameters can be categorized into three classes: the raw materials or substrate used (Tea and sucrose), the inoculum (SCOBY and kombucha tea), and the operational conditions (pH, temperature, and fermentation duration). Also, the database includes various tea types used for kombucha fermentation, such as black tea, green tea, waste tea, and the mixture of green and black tea in the ratio of 2:1. Some other tea varieties like corn silk tea, rooibos tea, oolong tea, white tea, red tea, and puerh tea are also being used widely in the regions of China (Chakravorty et al., 2019; Wu et al., 2013). In addition to these, operational parameters like the surface area and depth of culture vessels also influence the quality and quantity of cellulose produced (Abd El-Salam, 2012; Al-Kalifawi and Hassan, 2014; Ruka et al., 2012). However, they were not included in the study because of insufficient data. EDA reported the basic statistical interpretations on the distribution of data. Table 1 represents the statistical analysis of the input and output parameters used for machine learning model development. The count denotes the number of entries present in the database for a particular parameter. Mean represents the average value of each parameter. The outlook of the distribution of the data can be gathered from the four quartiles (25%, 50%, 75%, and 100%) and the standard deviation. In addition, minimum and maximum values showing the extreme values of the parameters are also presented.

Tea and sugar (sucrose) act as the primary nitrogen and carbon source required for the growth of bacteria and yeast. Greenwalt et al. (2000) claimed the sugared tea broth to be the best medium for improving biofilm yield. Yeast converts sucrose into monosaccharides (glucose and fructose) and they are oxidized to ethanol by alcoholic fermentation. Ethanol production in kombucha fermentation is considered to elevate the growth of acetic acid bacteria and stimulate the synthesis of bacterial cellulose (Soh and Lee, 2002). Caffeine present in tea leaves is believed to stimulate the bacterial cellulose synthesis (Chakravorty et al., 2019). The synthetic medium used for cellulose production in industries is Hestrin Schramm (HS) medium consisting of glucose, yeast extract, peptone and other additives. In comparison with HS medium, the sweetened tea broth is a low-cost medium with better production efficiency (Treviño-Garza et al., 2020). Moreover, sucrose is preferred over glucose because the increased production of gluconic acid from glucose leads to inhibition of cellulose synthesis (Greenwalt et al., 2000). The range of tea and sucrose concentrations in the database varies from 0–200 g/L and 0–550 g/L, respectively. This shows the extreme variable conditions under which the researches have been conducted. However, the optimum concentrations of tea and sucrose reported are between 5 and 10 g/L and 60–120 g/L, respectively (Laavanya et al., 2021). Malbaša et al. (2008) recorded the highest yield of SCOBY biofilm from 70 g/L sucrose concentration. The higher concentration of substrates is found to have a detrimental effect on the yield of biofilm over time (Al-Kalifawi and Hassan, 2014; Goh et al., 2012a). More than 75% of the datasets are observed to be in accordance with the reported optimal range of tea and sucrose concentrations.

Following the substrates, the inoculum also has an effect on biofilm formation. The inoculation is carried out by the addition of fresh SCOBY culture to the tea-sugar broth. SCOBY biofilm comprises a matrix of cellulose fibres that has microbial cells embedded in them. The culturing of the single strain of cellulose-producing bacteria is found to be less efficient than the SCOBY culture, considering cellulose production (Sharma and Bhardwaj, 2019). The fermented kombucha tea is also added to the culture broth. Though kombucha tea has microbial population, the key role is the reduction of pH to a slightly acidic level due to the presence of organic acids. Thus, contamination by pathogenic bacteria and fungi can be prevented, and the microbial population of SCOBY shall survive. Acetic acid is used as an alternative for kombucha tea in few pieces of literature (Santos Jr et al., 2009). The range of SCOBY and kombucha tea used for inoculation is 8–40 g/L and 5–25%, respectively.

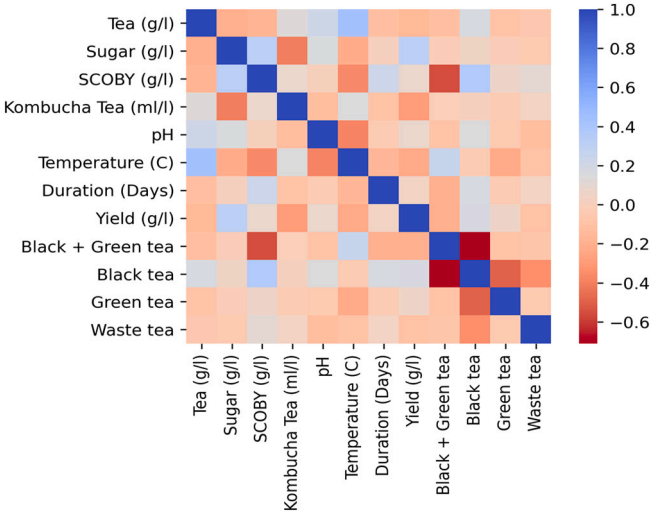

Fig. 1. Pearson correlation matrix of model parameters used for predicting SCOBY yield.

The operational conditions that are taken into consideration in this study are pH, temperature, and fermentation duration. According to the collected data, the pH lies in the range of 2–4, which is similar to the reported optimal values (Goh et al., 2012a). The pH lower than optimal value has a negative effect on the microbial growth and cellulose production, meanwhile the neutral pH supports the growth of contaminants (Sharma and Bhardwaj, 2019). Kombucha fermentation is usually performed at room temperature under dark conditions. The statistics of the database show the range of the temperature to be 20–45 °C. The period of fermentation can vary between 7 and 56 days, as observed from the database. However, a prolonged period of fermentation reduces cellulose's productivity for two major reasons: 1) The depletion of nutrients, and 2) The alcohol fermentation by yeast releases carbon dioxide (CO_2), which accumulates in the interface of the medium and the biofilm. In the long run, an anaerobic condition is created due to accumulation of CO_2 in the interface, which harms the microbes (Chen and Liu, 2000). During the course of fermentation, the pH of the medium further drops, indicating the occurrence of the fermentation process and the production of the organic acids (Ahmed et al., 2020; Muhialdin et al., 2019).

The wet weight of biofilm yield is in the range of 0.38–465 g/L, with an average of 55 g/L. The statistical analysis shows that over 75% yield lies between 13 and 64 g/L, similar to yield observed under normal fermentation conditions. In contrast, some literature has reported variation in the yield as a result of SCOBY with variation in the proportion of bacteria and yeast species and thus the cellulose production (Nguyen et al., 2008). The productivity ranges from 1 to 5 g/L·day, with an average of 4.7 g/L·day. Regarding tea type, the biofilm yield was relatively higher when green tea was used as a substrate, which is similar to the previous studies (Laavanya et al., 2021).

3.2. Correlation studies of input-output variables

To understand the relationship between the input-output parameters, Pearson correlation and pair plot have been analyzed. Fig. 1 represents the Pearson correlation plot, where the bar on the right indicates the scale of Pearson correlation coefficients from −1 to +1 that color coded as red to blue, respectively. The negative value signifies a negative linear correlation between the features, and the positive value implies the positive linear correlation between the features. The SCOBY and the mixture of green and black have a negative correlation. Almost all other parameters are found to have minimum correlations. The pair plot analysis showing a correlation between every two individual parameters used in this study. In addition, it helps understand the data

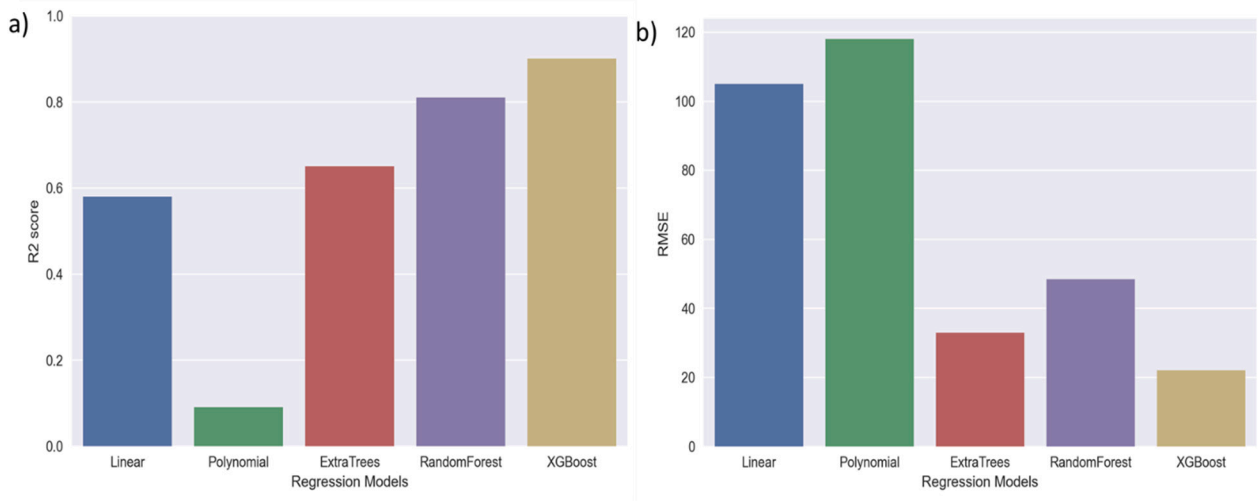

Fig. 2. Comparison of a) correlation coefficient (R^2) and b) root mean square error (RMSE) of regression models developed for SCOBY yield prediction.

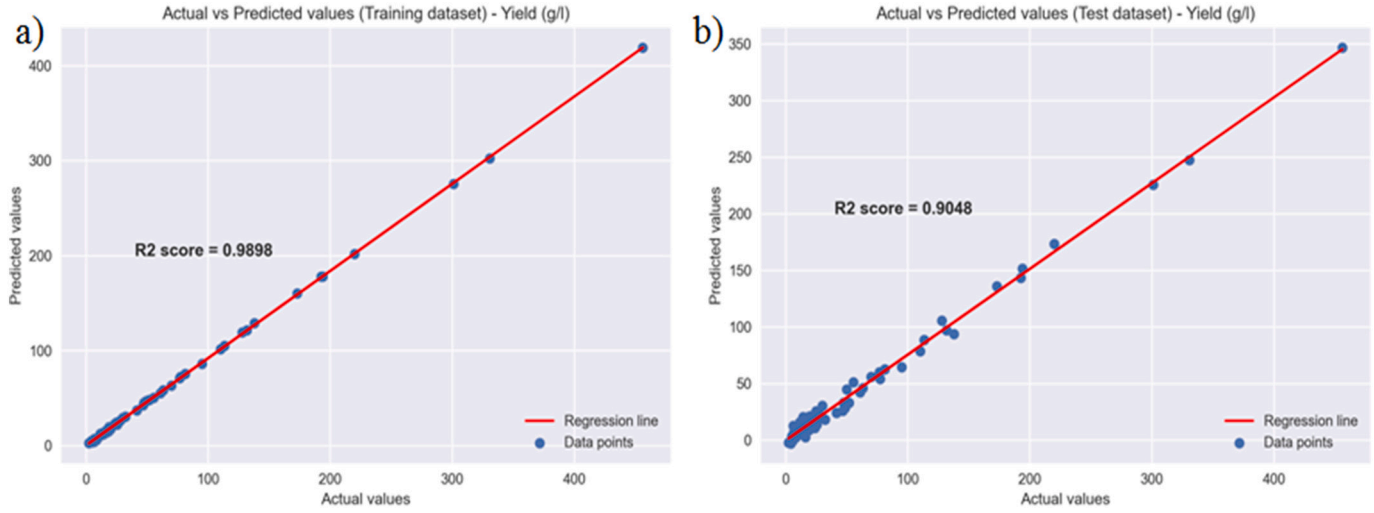

Fig. 3. Scatter plot for a) training and b) testing datasets used in XGB model.

distribution of a parameter in association with the other parameter. A better view of yield with influence of each parameter can be also be gathered by studying the pair plots.

3.3. Model selection and validation

In this study, the performance of the selected five regression models was evaluated using the R^2 and RMSE value. The five regression models were: Linear regression model, Polynomial regression model, ExtraTrees regression model, Random forest model, and XGB model.

The comparison chart for the RMSE values of the models is provided in the Fig. 2b. The models are shown in the descending order of RMSE values as, Polynomial > Linear > Random forest > ExtraTrees > XGB. The results show that the relationship between the input and target variables is not linear. XGB outperformed all other regression models with the lowest RMSE value. In addition to RMSE, the R^2 value also assists in the selection of a better model. The models with low to high R^2 values are, Polynomial < Linear < ExtraTrees < Random forest < XGB (Fig. 2a). Clearly, XGB is the most accurate model with the highest R^2 and lowest RMSE value; hence, selected for further studies.

3.4. XGB model development

For the development of the robust XGB model, the concern of data overfitting needs to be resolved with hyperparameter tuning and cross-validation. *k*-Fold cross-validation is a reliable method to validate the performance of the model with the limited datasets that are available. The entire dataset was made into *k* subsets, and *k-1* subsets were used for training. The model was validated with the last subset, and in a similar way, 10-fold cross-validation was performed. After validation, the model was trained and tested further. The datasets were randomly split into training and testing datasets in the ratio of 80:20. A learning curve that indicates the reduction in RMSE value with an increase in the number of iterations was studied and found that the curve stabilized at around 140 iterations. Scatter plots of actual vs. predicted values for training and testing datasets were generated (Fig. 3), and the R^2 values were 0.9898 and 0.9048, respectively.

3.5. Studying the significance of input variables on SCOBY yield

Feature importance indicates the significance of each parameter on the model output. Based on the feature importance, all the input parameters can be ranked from the most influential to the least significant factors that affect the SCOBY yield. From Fig. 4, it is evident that

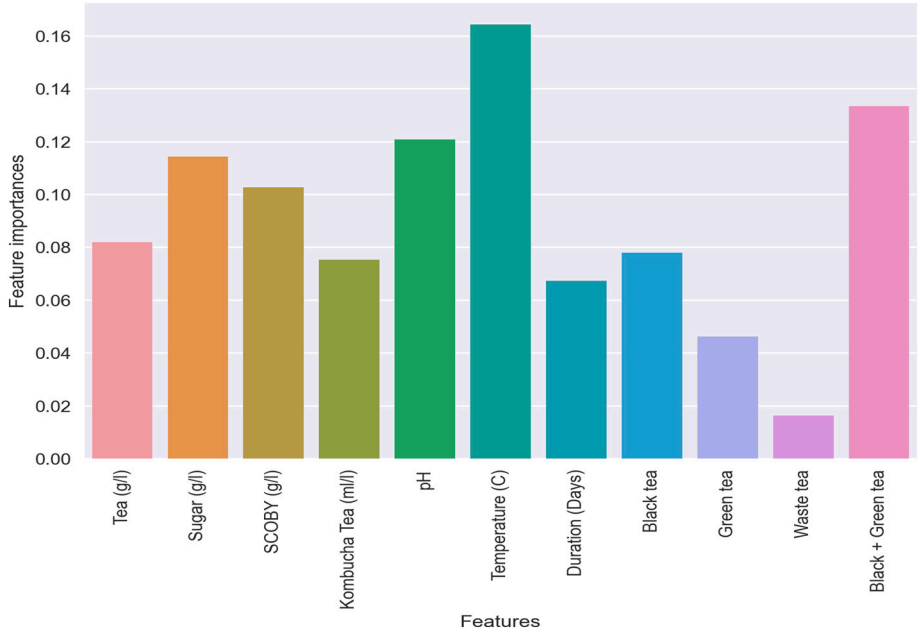

Fig. 4. Feature importance plot of influence of input variables on SCOBY yield.

temperature has the most impact on the model output with F score (feature importance) greater than 0.16. Following that, black + green tea, sugar concentration, pH, and SCOBY concentration are among the top five influential factors. It should be noted that the influence may be positive or negative on the output, which can be studied from Shapley values. Since the mixture of black and green tea belongs to the category of type of tea; it would be appropriate to compare it with the other types and interpreted separately. Therefore, among the other four types of tea, the mixture of black + green tea (2:1 ratio) and waste tea had the highest and lowest influence on the SCOBY yield, respectively.

De Filippis et al. (2018) reported temperature of kombucha fermentation supports the growth of specific bacterial species of SCOBY. Thus, the species selection trait of temperature could eventually influence the yield of bacterial cellulose. Though SCOBY has diverse microbial population, the proportion of growth and its metabolic activity depends on the environmental conditions of the culture. This may be the underlying reason for temperature being the most influential parameter in the study.

After temperature, pH was the most influential factor in the production of SCOBY cellulose according to the importance plot. It acts as an indicator for optimum fermentation conditions to obtain good quality kombucha tea and high SCOBY yield. The pH of the culture varies throughout the period of fermentation. Bacteria and yeast present in the inoculum grows and produces organic metabolites such as acetic acid, which consecutively reduces the pH. Also, acetic acid maintains the oxidation-reduction balance during fermentation (Laureys et al., 2020; Laavanya et al., 2021). But higher acetic acid production lowers the pH and below a certain level it might cause acid stress to bacteria and reduce the SCOBY production.

3.6. Combined influence of input parameters on SCOBY yield

SHAP dependence plots can emphasize the interactive effects of the input parameters on the model output. In each plot, the effect of a primary factor on the yield is presented along with its closely associated secondary variable. The x axis represents the primary input factor and the y axis represents its corresponding SHAP value. The secondary input variable is represented by the color bar on the y axis, ranging from high to low. From these plots (Fig. 5), the combination of process conditions that has positive impact on the output could be deducted and optimized for the production processes.

It can be seen that low concentration of sugar has a negative impact on the yield irrespective of the tea concentration. With increase in concentration of sugar to the range of 70–100 g/L and with low concentration of tea, the SHAP value of sugar increased positively and found to be the favorable condition for attaining maximum yield. This observation was similar to the optimum conditions reported in the literatures (Abd El-Salam, 2012; Laavanya et al., 2021). The use of sugar less than the optimum concentration can reduce the yield, as high sugar reserves will be consumed for the growth of bacteria leading to exhaustion of substrate. Additionally, the mid to high values of sugar concentration had negative effect on the yield of the output during the first week of fermentation. The productivity of bacterial cellulose will be comparatively lower during the initial days of fermentation, as it takes time for the bacteria to stabilize and increase in population.

The key observations from the SHAP dependence plots that had positive effect on the yield of bacteria cellulose are: temperature in the range of 24–28 °C, low concentration of tea (5–10 g/L), SCOBY concentration (30 g/L), 10–14 days of fermentation, and slightly acidic pH (3.5–4.5). These conditions can be conducive for increasing the biofilm yield.

Tea leaves are rich in polyphenols that are known for antioxidant and antibacterial activity as well. So, higher concentrations of tea might have inhibition effect on growth of bacteria and responsible for low SCOBY biofilm yield (Sharma and Bhardwaj, 2019). In few studies, the addition of kombucha tea was replaced with acetic acid. Thus, a positive impact can be observed even without the addition of kombucha tea inoculum. Yet, the optimum concentration was found to be 100 ml/L. Further, increase in concentration of kombucha tea (>100 ml/L) would ultimately the affect the yield of biofilm, since the pH will become highly acidic and unsuitable for the growth of yeast and bacteria. On the contrast, Soh and Lee (2002) reported that fermentation with an initial pH in the range of 2.5 has lower productivity initially but showed increase in yield on long run. This ultimately depends on the microbial diversity of the SCOBY, as it may have high acid tolerant strains of bacteria. Kombucha fermentation with an initial pH above 4 yielded higher SCOBY biofilm during early days. In case of tea type, the use of green tea had relatively higher positive impact followed by black tea. However, the mixture of black and green tea was observed to have negative impact on the yield. The reason for this effect is unclear and

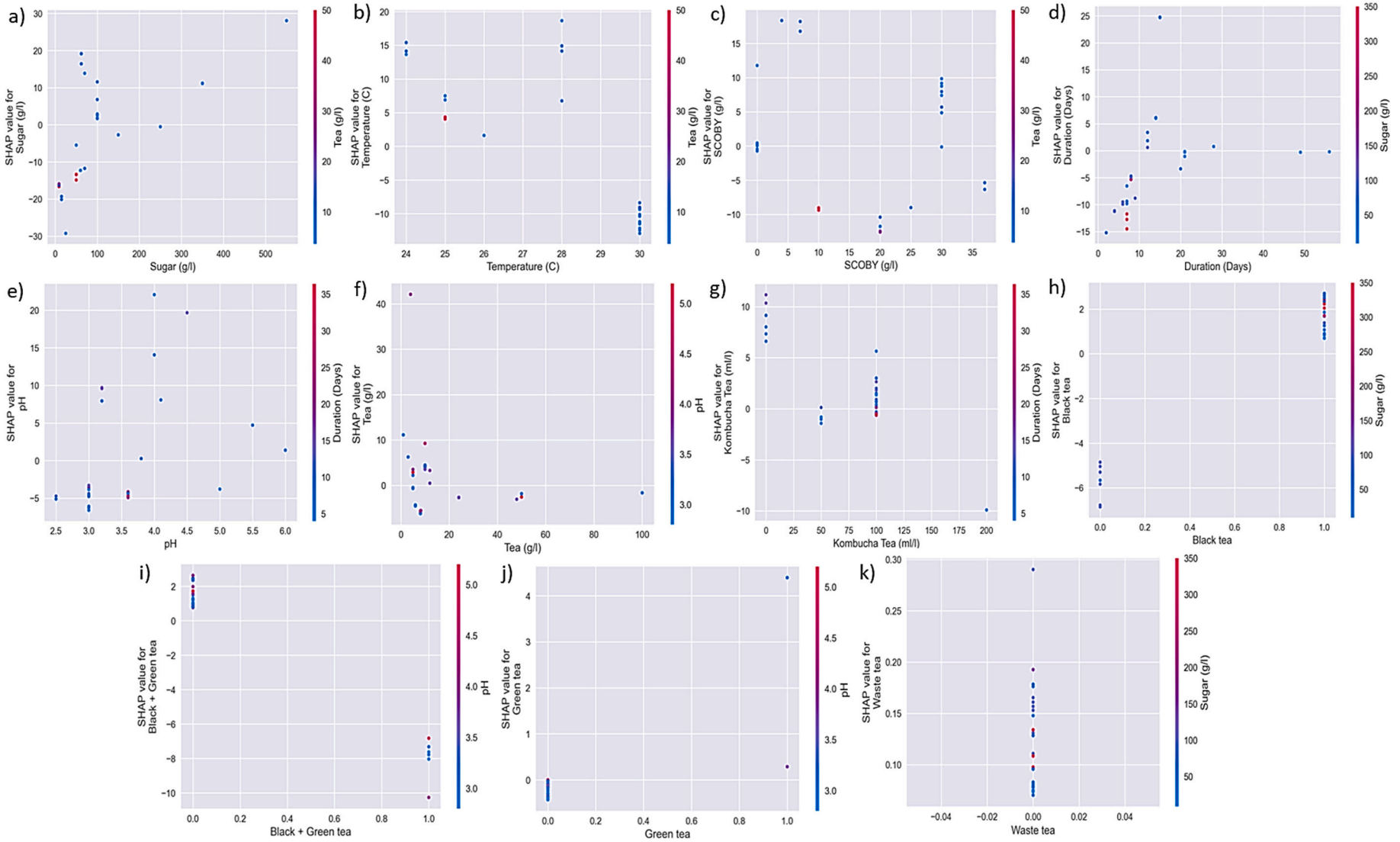

Fig. 5. SHAP dependence plots for studying the effect input variables on SCOBY yield.

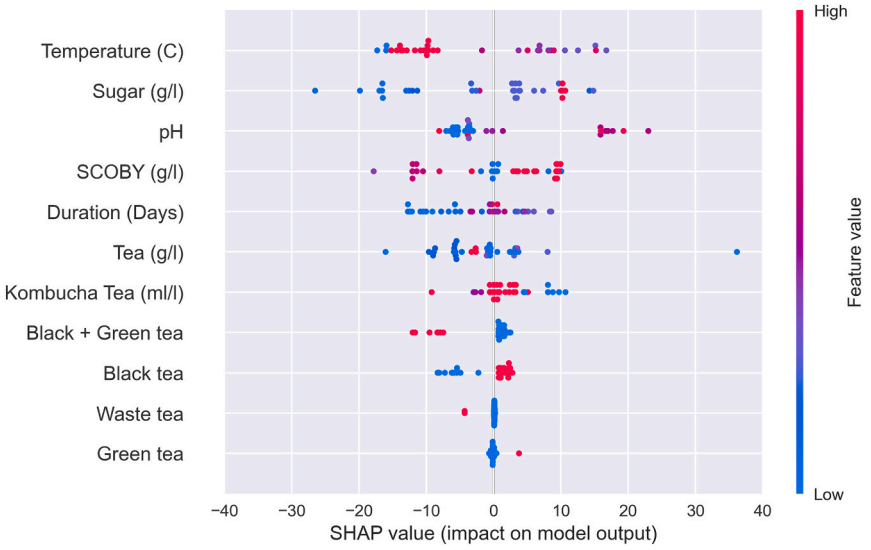

Fig. 6. Summary plot for understanding the impact of input variables on SCOBY yield.

requires thorough analysis of tea components. The use of waste tea had almost neutral impact on the yield output.

A summary plot is the combination of feature importance plot and SHAP plots. From Fig. 6, the effect of each on the model output can be elucidated. According to the summary plot, sugar concentration had the most impact with high and mid-range sugar concentrations having a positive effect on yield. In contrast, high amount of sugar can also affect the yield of bacterial cellulose due to product inhibition reactions (Al-Kalifawi and Hassan, 2014). The increase of metabolic activity with high amounts of sugar, trigger higher production of gluconic acid, which consecutively ceases the bacterial cellulose synthesis (Greenwalt et al., 2000; Sharma and Bhardwaj, 2019). Another possible reason is that the imbalance in nutrient uptake and utilization rate can lead to the accumulation of substances inside the cell (Goh et al., 2012b; Sharma and Bhardwaj, 2019).

Similar to literature evidences, low to mid-range of temperature affect the model output constructively, thus maximizing the yield. A key observation from the summary plot was that very high and low duration of fermentation had affected the yield of cellulose negatively. According to Gargey et al. (2019), the reduction in productivity of bacterial cellulose after a period of two weeks might be because of the microbes of SCOBY entering the stationary phase. All the literatures used for data collection have practiced static fermentation experiments. In such conditions, bacteria deplete the dissolved oxygen content within two weeks' time period of fermentation. Therefore, only the bacteria attached to SCOBY on the air-liquid interface can actively produce cellulose and the bacteria present in the broth has low production capability (Al-Kalifawi, 2014). This explains the negative impact on SCOBY yield by longer duration of fermentation. Tea type had lowest impact on the output compared to the other parameters. Black tea and green tea if used as substrate, had positive impact on the model. Also, in accordance to the reports, a slightly acidic pH was found to benefit the yield of the bacterial cellulose. Acetic acid bacteria count increased consistently under mild acidic conditions and as a result production efficiency of bacterial cellulose is improved (De Filippis et al., 2018). The effect of SCOBY inoculum could not be deduced accurately from the summary plot. However, even low concentration of kombucha tea inoculum affects the yield in a positive manner. Studies suggested that the SCOBY biofilm production has inverse relationship with quantity of kombucha tea broth used (Soh and Lee, 2002).

4. Conclusion

XGB model was developed to comprehend the influencing factors on SCOBY formation in Kombucha tea fermentation that has 90.48% similarity between the actual and predicted values. Importance and SHAP dependence plots revealed that fermentation temperature was the most significant parameter followed by sugar concentration and pH. Green tea can increase the production of SCOBY cellulose. The optimal conditions for higher yield could be gathered from Summary plot analysis. The inclusion of other highly influential input parameters such as surface area of fermenter and SCOBY type in forthcoming model development could make the predictions much more reliable and assist in scaling-up.

CRediT authorship contribution statement

Conceptualization: Balasubramanian Paramasivan, *Methodology*: Vimaladasan Senthamizhan, *Data collection*: Priyadharshini Thangaraj, *Formal analysis and investigation*: Priyadharshini Thangaraj, Nageshwari Krishnamoorthy, Vimaladasan Senthamizhan, *Writing – original draft preparation*: Priyadharshini Thangaraj, Nageshwari Krishnamoorthy, *Writing – review and editing*: Balasubramanian Paramasivan, Nageshwari Krishnamoorthy, *Funding acquisition*: Balasubramanian Paramasivan, Parag Prakash Sutar, *Resources & Supervision*: Balasubramanian Paramasivan.

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.

Acknowledgements

The authors acknowledge the Department of Biotechnology (DBT) of Government of India for funding the research under Biotechnology Ignition Grant (BIG) of Biotechnology Industry Research Assistance Council (BIRAC) [BIRAC/KIIT0471/BIG-13/18]. The authors thank the Department of Biotechnology and Medical Engineering of National Institute of Technology Rourkela for providing the research facility. The authors greatly acknowledge the guidance and mentorship from [Late] Prof. Rasu Jayabalan of National Institute of Technology Rourkela.

Appendix A. Supplementary data

All data generated or analyzed during this study are included in this published article (and its Supplementary information files). Supplementary data to this article can be found online at https://doi.org/10.1016/j.biteb.2022.101027.

References

Abd El-Salam, S.S., 2012. Bacterial cellulose of kombucha mushroom tea. N. Y. Sci J. 5 (4), 81–87.

Ahmed, R.F., Hikal, M.S., Abou-Taleb, K.A., 2020. Biological, chemical and antioxidant activities of different types Kombucha. Ann. Agric. Sci. 65 (1), 35–41.

Al-Kalifawi, E.J., 2014. Produce bacterial cellulose of kombucha (Khubdat Humza) from honey. J. Genet. Environ. Resour. 2 (1), 39–45.

Al-Kalifawi, E.J., Hassan, I.A., 2014. Factors Influence on the yield of bacterial cellulose of Kombucha (Khubdat Humza). Baghdad Sci. J. 11 (3), 1420–1428.

Balkanski, E., Rubinstein, A., Singer, Y., 2017. The limitations of optimization from samples. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1016–1027.

Chakravorty, S., Bhattacharya, S., Bhattacharya, D., Sarkar, S., Gachhui, R., 2019. Kombucha: a promising functional beverage prepared from tea. In: Non-alcoholic Beverages. Woodhead Publishing, pp. 285–327.

Chen, C., Liu, B.Y., 2000. Changes in major components of tea fungus metabolites during prolonged fermentation. J. Appl. Microbiol. 89 (5), 834–839.

De Clercq, D., Wen, Z., Fei, F., Caicedo, L., Yuan, K., Shang, R., 2020. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Sci. Total Environ. 712, 134574.

De Filippis, F., Troise, A.D., Vitaglione, P., Ercolini, D., 2018. Different temperatures select distinctive acetic acid bacteria species and promotes organic acids production during Kombucha tea fermentation. Food Microbiol. 73, 11–16.

de Oliveira Barud, H.G., da Silva, R.R., da Silva Barud, H., Tercjak, A., Gutierrez, J., Lustri, W.R., de Oliveira Junior, O.B., Ribeiro, S.J., 2016. A multipurpose natural and renewable polymer in medical applications: Bacterial cellulose. Carbohydr. Polym. 153, 406–420.

Gargey, I.A., Indira, D., Jayabalan, R., Balasubramanian, P., 2019. Optimization of etherification reactions for recycling of tea fungal biomass waste into carboxymethylcellulose. In: Green Buildings and Sustainable Engineering. Springer, Singapore, pp. 337–346.

Goh, W.N., Rosma, A., Kaur, B., Fazilah, A., Karim, A.A., Bhat, R., 2012a. Fermentation of black tea broth (Kombucha): I. Effects of sucrose concentration and fermentation time on the yield of microbial cellulose. Int. Food Res. J. 19 (1), 109–117.

Goh, W.N., Rosma, A., Kaur, B., Fazilah, A., Karim, A.A., Bhat, R., 2012b. Microstructure and physical properties of microbial cellulose produced during fermentation of black tea broth (Kombucha)II. Int. Food Res. J. 19 (1), 153–158.

Goodswen, S.J., Barratt, J.L., Kennedy, P.J., Kaufer, A., Calarco, L., Ellis, J.T., 2021. Machine learning and applications in microbiology. FEMS Microbiol. Rev. 45 (5), 1–19.

Greenwalt, C.J., Steinkraus, K.H., Ledford, R.A., 2000. Kombucha, the fermented tea: microbiology, composition, and claimed health effects. J. Food Prot. 63 (7), 976–981.

Jayabalan, R., Marimuthu, S., Swaminathan, K., 2007. Changes in content of organic acids and tea polyphenols during kombucha tea fermentation. Food Chem. 102 (1), 392–398.

Jayabalan, R., Malini, K., Sathishkumar, M., Swaminathan, K., Yun, S.E., 2010. Biochemical characteristics of tea fungus produced during kombucha fermentation. Food Sci. Biotechnol. 19 (3), 843–847.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.

Kamiński, K., Jarosz, M., Grudzień, J., Pawlik, J., Zastawnik, F., Pandyra, P., Kołodziejczyk, A.M., 2020. Hydrogel bacterial cellulose: a path to improved materials for new eco-friendly textiles. Cellulose 27 (9), 5353–5365.

Laavanya, D., Shirkole, S., Balasubramanian, P., 2021. Current challenges, applications and future perspectives of SCOBY cellulose of Kombucha fermentation. J. Clean. Prod. 295, 126454.

Laberge, Y., 2011. Advising on research methods: a consultant's companion. J. Appl. Stat. 38 (12), 2991.

Laureys, D., Britton, S.J., De Clippeleer, J., 2020. Kombucha tea fermentation: a review. J. Am. Soc. Brew. Chem. 78 (3), 165–174.

Lee, T.H., Ullah, A., Wang, R., 2020. Bootstrap aggregating and random forest. In: Macroeconomic Forecasting in the Era of Big Data. Springer, Cham, pp. 389–429.

Liu, J., Li, Q., Chen, W., Yan, Y., Qiu, Y., Cao, T., 2019. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. Int. J. Hydrog. Energy 44 (11), 5470–5480.

Malbaša, R., Lončar, E., Djurić, M., 2008. Comparison of the products of Kombucha fermentation on sucrose and molasses. Food Chem. 106 (3), 1039–1045.

Markov, S.L., Malbaša, R.V., Hauk, M.J., Cvetković, D.D., 2001. Investigation of tea fungus microbe assotiations: I: the yeasts. Acta Period. Technol. 32, 133–138.

Maulud, D., Abdulazeez, A.M., 2020. A Review on Linear Regression Comprehensive in Machine learning. J. Appl. Sci. Technol. Trends 1 (4), 140–147.

Muhialdin, B.J., Osman, F.A., Muhamad, R., Che Wan Sapawi, C.W.N.S., Anzian, A., Voon, W.W.Y., Hussin, A.S., 2019. Effects of sugar sources and fermentation time on the properties of tea fungus (kombucha) beverage. Int. Food Res. J. 26 (2), 481–487.

Mukadam, T.A., Punjabi, K., Deshpande, S.D., Vaidya, S.P., Chowdhary, A.S., 2016. Isolation and characterization of bacteria and yeast from Kombucha tea. Int. J. Curr. Microbiol. 5 (6), 32–41.

Naomi, R., Bt Hj Idrus, R., Fauzi, M.B., 2020. Plant- vs. Bacterial-derived cellulose for wound healing: A review. Int. J. Environ. Res. Public Health 17 (18), 6803.

Nguyen, V.T., Flanagan, B., Gidley, M.J., Dykes, G.A., 2008. Characterization of cellulose production by a Gluconacetobacter xylinus strain from Kombucha. Curr. Microbiol. 57 (5), 449–453.

Pathy, A., Meher, S., Balasubramanian, P., 2020. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Res. 50, 102006.

Ruka, D.R., Simon, G.P., Dean, K.M., 2012. Altering the growth conditions of Gluconacetobacter xylinus to maximize the yield of bacterial cellulose. Carbohydr. Polym. 89 (2), 613–622.

Santos Jr., R.J., Batista, R.A., Rodrigues Filho, S.A., Lima, A.S., 2009. Antimicrobial activity of broth fermented with kombucha colonies. J. Microbial. Biochem. Technol. 1 (1), 72–78.

Sharma, C., Bhardwaj, N.K., 2019. Biotransformation of fermented black tea into bacterial nanocellulose via symbiotic interplay of microorganisms. Int. J. Biol. Macromol. 132, 166–177.

Soh, H.S., Lee, S.P., 2002. Production of microbial cellulose and acids in kombucha. Prev. Nutr. Food Sci. 7 (1), 37–42.

Treviño-Garza, M.Z., Guerrero-Medina, A.S., González-Sánchez, R.A., García-Gómez, C., Guzmán-Velasco, A., Báez-González, J.G., Márquez-Reyes, J.M., 2020. Production of Microbial Cellulose Films from Green Tea (Camellia Sinensis) Kombucha with various carbon sources. Coatings 10 (11), 1132.

Volk, M.J., Lourentzou, I., Mishra, S., Vo, L.T., Zhai, C., Zhao, H., 2020. Biosystems design by machine learning. ACS Synth. Biol. 9 (7), 1514–1533.

Wang, J., Xu, J., Zhao, C., Peng, Y., Wang, H., 2019. An ensemble feature selection method for high-dimensional data based on sort aggregation. Syst. Sci. Control. Eng. 7 (2), 32–39.

Wang, Y., Huntington, T., Scown, C.D., 2021. Tree-based automated machine learning to predict biogas production for anaerobic co-digestion of organic waste. ACS Sustain. Chem. Eng. 9 (38), 12990–13000.

Weichert, D., Link, P., Stoll, A., Rüping, S., Ihlenfeldt, S., Wrobel, S., 2019. A review of machine learning for the optimization of production processes. Int. J. Adv. Manuf. Syst. 104 (5), 1889–1902.

Wu, Y., Chen, Q., Ruan, H., He, G., 2013. Optimization of liquid fermentation process for improved exo-polysaccharides production by kombucha ZJU1. Adv. J. Food Sci. Technol. 5 (2), 217–224.