Biofuel Research Journal 38 (2023) 1859-1875

Biofuel Research Journal

Journal homepage: www.biofueljournal.com

Review Paper

A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route

Ahmet Coşgun[1], M. Erdem Günay[2], Ramazan Yıldırım[1,*]

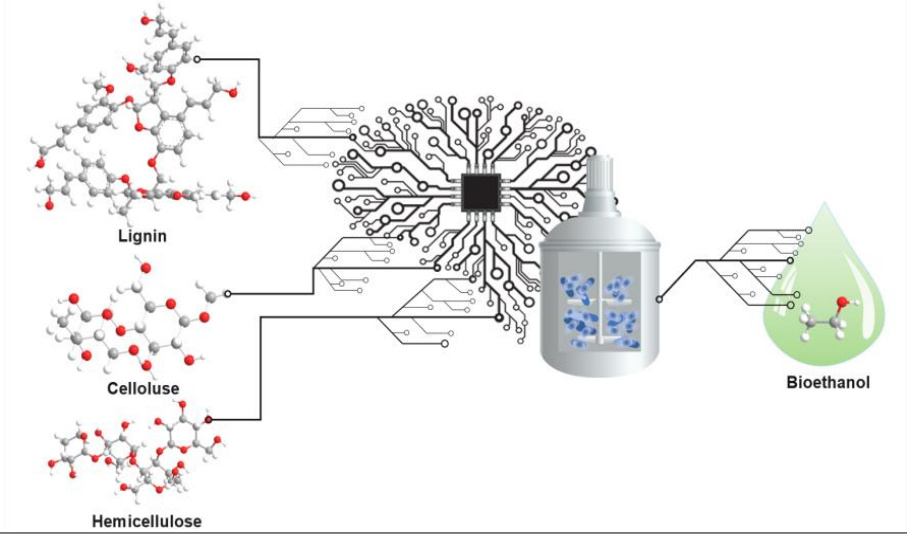
[1] *Department of Chemical Engineering, Boğaziçi University, 34342, Bebek-Istanbul, Turkey.*

[2]*Department of Energy Systems Engineering, Istanbul Bilgi University, 34060, Eyup-Istanbul, Turkey.*

HIGHLIGHTS

- Studies on machine learning (ML) applications for lignocellulosic ethanol production are critically reviewed.
- Bibliometric research and future perspectives on ML applications are provided.
- ANN is the most commonly used algorithm (appearing in almost 90% of articles).
- Bioethanol concentration is the most common output variable in the fermentation step.
- Fermentable sugar and glucose concentration are studied most in studies focused on the hydrolysis step.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 March 2023
Received in revised form 5 May 2023
Accepted 8 May 2023
Published 1 June 2023

Keywords:
Biofuel production
Bioethanol
2nd generation feedstock
Lignocellulosic ethanol
Cellulosic ethanol
Machine learning

ABSTRACT

In this work, machine learning (ML) applications in lignocellulosic bioethanol production were reviewed. First, the pretreatment-hydrolysis-fermentation route, the most commonly studied alternative, was summarized. Next, a bibliometric analysis was performed to identify the current trends in the field; it was found that ML applications in the field are not only increasing but also expanding their relative share in publications, with bioethanol seeming to be the most frequently researched topic while biochar and biogas are also receiving increased attention in recent years. Then, the implementation of ML for lignocellulosic bioethanol production *via* this route was reviewed in depth. It was observed that artificial neural network (ANN) is the most commonly used algorithm (appeared in almost 90% of articles), followed by response surface methodology (RSM) (in about 25% of articles) and random forest (RF) (in about 10% of articles). *Bioethanol concentration* is the most common output variable in the fermentation step, while *fermentable sugar* and *glucose concentration* are studied most in hydrolysis. The datasets are usually small, while the fitnesses of the models (R^2) are usually high in the papers reviewed. Finally, a perspective for future studies, mostly considering improving data availability, was provided.

* Corresponding author at:
E-mail address: yildirra@boun.edu.tr

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Contents

Abbreviations

AFEX	Ammonia fiber expansion
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
ANN-PSO	Artificial neural network in combination with particle swarm optimization
CBF	Consolidated bioprocessing
CEF	Cellulose enrichment factor
CSF	Combined severity factor
DA	Dilute-acid treatment
DES	Deep eutectic solvent
DFT	Density functional theory
DT	Decision tree
DTR	Decision tree regression
FUZZY-GAP	Fuzzy system
GB	Gradient boosting
HTC	Hydrothermal carbonization
HTL	Hydrothermal liquefaction
IL	Ionic liquids
LCB	Lignocellulosic biomass
LCW	Lignocellulosic waste
MARS	Multivariate adaptive regression splines
MSW	Municipal solid waste
MW	Microwave
OD-MS	Optimized decision-making system
OV	Organosolv
PRM	Polynomial regression model
RBF-PSO	Radial basis functions in combination with particle swarm optimization
RF	Random forest
RSM	Response surface methodology
SCB	Sugarcane bagasse
SE	Saccharification efficiency
SHF	Separate hydrolysis and fermentation
SR	Solid recovery
SSCF	Simultaneous saccharification and co-fermentation
SSF	Simultaneous saccharification and fermentation
STE	Steam explosion
SVM	Support vector machine
US	Ultrasound

1. Introduction

Biofuel production from biomass, with negligible negative impacts on the environment due to the fast bioenergy cycle, is a sustainable way to meet the growing fuel demand. Among biofuels, bioethanol can be an alternative fuel for gasoline-powered vehicles. Bioethanol can be produced from sugars (*e.g.*, glucose) requiring fermentation followed by distillation and further purification steps, or starch needing hydrolysis as an additional step before fermentation (**Fig. 1**) (Brandt et al., 2013). Although converting simple sugars or starch-based biomass to ethanol through fermentation is quite straightforward, these first-generation feedstocks are no longer preferred because of the large amount of land required for agriculture, creating undesired competition with the food chain. On the other hand, lignocellulosic biomass (LCB) is not directly tied to food production and is readily available worldwide as agricultural wastes or forest residues (Kumar et al., 2016; Liu et al., 2019). Moreover, it is the most plentiful raw material on Earth that can be used to produce biofuels.

Different technological pathways exist to produce ethanol from LCB, as shown in **Figure 2** and as discussed in detail elsewhere (Aui et al., 2021). Among these methods, the gasification process is carried out at high temperatures with partial air, producing liquid hydrocarbons, biochar, and syngas as the main products; the syngas can also be fermented to ethanol *via* anaerobic digestion (Griffin and Schultz, 2012). Pyrolysis and liquefaction are also among the methods used to break down biomass into useful products (*e.g.*, bio-oil). In addition, the sugar content in the bio-oil can also be extracted and fermented to make bioethanol (Li et al., 2018). Nevertheless, direct fermentation is the predominant technological pathway for bioethanol production (**Fig. 1**), and this work focuses on this pathway.

Due to the complexity of the process variables associated with the conversion of LCB to ethanol, machine learning (ML) can help to determine the optimal experimental conditions leading to the highest conversion with the most feasible combination of variables. ML is a field of artificial intelligence that attempts to create and enhance computer programs that can automatically learn from past data using various algorithms. These programs can be used to study datasets, identify previously unknown trends and patterns, generalize data, develop models, or derive heuristics-based rules (Erdem Günay and Yıldırım, 2020); they are exceptionally good at detecting nonlinear correlations between input and output variables (Bannor and Acheampong, 2019). Although numerous ML algorithms are available and their numbers are continuously increasing, they mostly accomplish some specific tasks such as *prediction* or *classification* of the outcome from a new set of descriptors, *clustering* of the data based on the similarities of descriptors, or *associating/correlating* descriptors with each other and also with the output variables (Larose and Larose, 2014).

Typically, different algorithms are used for various tasks though some algorithms can be used for more than one task, and some tasks can be accomplished by multiple algorithms. For example, artificial neural networks (ANN), support vector machines, and random forest regression can be used for *prediction*, while decision/regression trees and support vector machines can be used for both regression and classification. Although specific tasks and algorithms require specific attention, the implementation of ML can be summarized in three main steps: *constructing the dataset* (*e.g.*, collecting data from experimental works in literature), *selecting and implementing ML algorithms* (*e.g.*, selecting ANNs to predict the yield, optimizing model hyperparameters and validating the model performance), and *interpreting the results* (*e.g.*, understanding the effects and significance of descriptors) (Alpaydin, 2020). Each step is important for the successful implementation of ML.

Indeed, ML has grown significantly over the past few years in various areas, including energy and fuels. There are also very successful applications of ML in biofuel research, as outlined by some reviews covering the entire field (Wang et al., 2022) or specific studies such as ANN

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

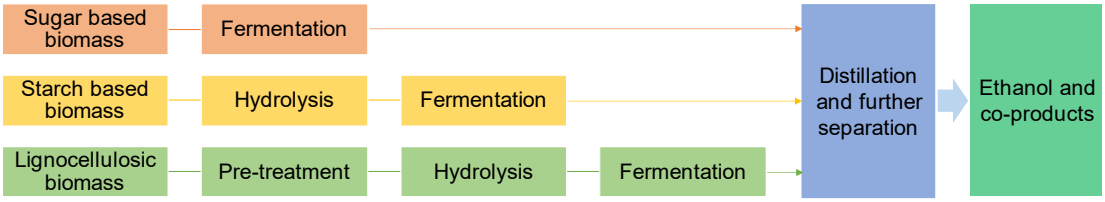


Fig. 1. Fermentation-based transformation of different feedstocks into ethanol.

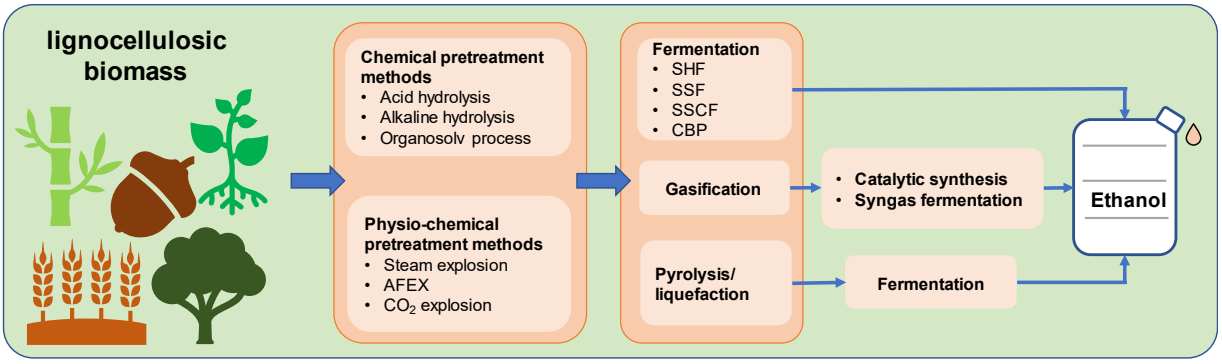


Fig. 2. Cellulosic ethanol conversion pathways and typical pretreatment methods (adapted from Aui et al. (2021)).

applications (Sewsynker-Sukai et al., 2016; Pradhan et al., 2022;), or reviews of ML for gaseous biofuels (Kucharska et al., 2018). Applications of ML for LCB have also been reviewed by a variety of research groups (Li et al., 2022; Li et al., 2023). Reviews concentrating on waste treatment (Guo et al., 2021), pyrolysis (Ge et al., 2021), and LCB pretreatment (Xu and Huang, 2014) are some of those examples.

Table 1 summarizes the reviews on bioethanol production through fermentative pathways (including the present paper) to illustrate each work's contribution to the field and show the gap that may be filled with the present study. As the table indicates, the present work can be differentiated from the others in two respects. First, the previously published literature either concentrated on specific value chain steps or specific ML algorithms, whereas our work covered all ML applications in all steps of bioethanol production *via* the fermentation route. Second, we performed an exploratory analysis through literature so that the shifting trends could also be seen to put the reviewed material in time perspective. To the best of our knowledge, there are no review papers with such coverage. In light of these, first, we summarized the lignocellulosic bioethanol production process through the pretreatment-hydrolysis-fermentation route, followed by an extensive text mining analysis. Next, the manuscript reviews and evaluates ML utilization in the field and, finally, provides a comprehensive perspective for future applications.

2. Lignocellulosic biomass and its conversion to bioethanol *via* fermentation

Compared to sugar and starch-based feedstock, LCB is more complex, and understanding its structure, especially at a molecular level, is critical (Liu et al., 2019). It mainly consists of cellulose, hemicellulose, and lignin; all of which are tangled up in one another to make lignin-carbohydrate complexes (**Fig. 3**). Cellulose is the primary component of plant cell walls giving them rigidity and strength, and it is the largest carbohydrate in the LCB accounting for 40-60% of the weight. Hemicelluloses are the second most abundant carbohydrate in LCB (about 20-35% of weight); they are heterogeneous polysaccharides, including several hexose sugars (*e.g.*, glucose, mannose, and galactose) and pentose sugars (*e.g.*, arabinose and xylose) (Brandt et al., 2013). Finally, as the remaining part, lignin is an aromatic, water-insoluble polymer that provides the plant with water-proofing ability, structural strength, and resilience (Zoghlami and Paes, 2019). Compared to cellulose and hemicellulose, lignin (the protective structure) is especially resistant to biological breakdown.

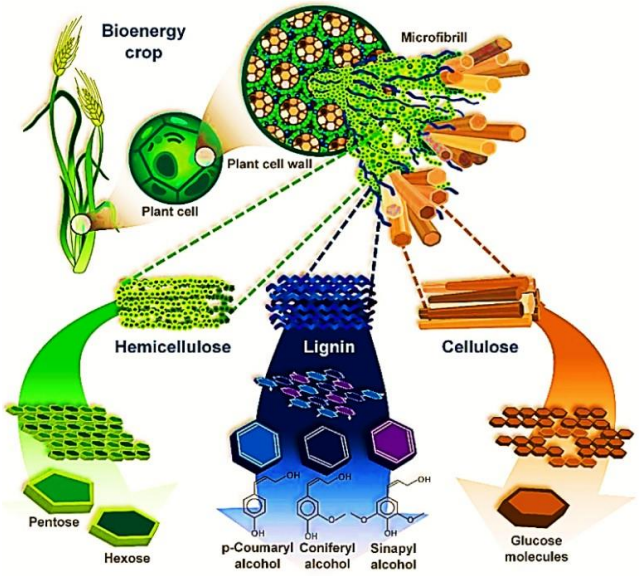


Fig. 3. Structure of lignocellulosic biomass and its main constituents. Reproduced from Ponnuchamy (2022) with permission from Elsevier, Copyright© 2022.

Unfortunately, due to the complex nature of the LCB, conversion to ethanol is not straightforward and highly complicated. The fermentation pathway can be carried out in three steps; pretreatment, hydrolysis, and fermentation. The components of LCB are bonded with strong covalent bonds, Van der Wall's forces, and various intermolecular bridges forming a strong and complex structure that is highly stable against hydrolysis

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Table 1.
Comparison of the coverage of the reviews previously published on ML-aided bioethanol production *via* fermentation route and the present work.

Bibliometric analysis	ML method [a]	Feedstock	Pretreatment	Hydrolysis	Fermentation	Reference
✗	all	all	✗	✗	✓	Wang et al. (2022)[b]
✗	ANN, ANFIS	all	✓	✓	✓	Pradhan et al. (2022)
✗	all	all	✗	✗	✓	Sewsynker-Sukai et al. (2016)[b]
✗	all	waste	✗	✗	✓	Li et al. (2022)
✓	all	all	✗	✗	✓	Culaba et al. (2022)[b]
✓	all	all	✓	✓	✓	**This Review**

✓: Included.
✗: Not included.
[a] ANN: artificial neural network; ANFIS: adaptive neuro-fuzzy inference system.
[b] Includes other biofuels.

(Kumar et al., 2010). Hence, pretreatment is required to separate the lignin and recover cellulose and hemicellulose for conversion to ethanol (Cheah et al., 2020). The separated lignin can be combusted to generate heat or sold in the market (Aui et al., 2021). The remaining complex polymer structures (*i.e.*, cellulose and hemicellulose) are converted into simple sugar molecules *via* hydrolysis. Then, in the final step, the fermentable simple sugar molecules are converted into ethanol (Charte et al., 2017). The pretreatment of LCB is the most expensive process in the pathway, accounting for approximately 20% of the total cost (Yang and Wyman, 2008); the hydrolysis and fermentation steps are also not easy due to the presence of a complex mixture of different sugars, which has a detrimental impact on the economic feasibility of the process.

2.1. Pretreatment

Several pretreatment processes exist, such as physical/physicochemical, chemical, and biological treatments. Mechanical size reduction, such as chopping, is one of the physical methods for increasing the surface of LCB; the ultra-fine milling process can also be used to reduce cellulose polymerization and crystallinity, although it is costly (Liu et al., 2019). Microwave heating is another potential alternative pretreatment for lignocellulosic materials as it eliminates the need for solvents, separating agents, and other auxiliary chemicals, it produces no smoke or waste and reduces the processing time and energy compared to other heating systems (Aguilar-Reynosa et al., 2017). Additionally, liquid hot water pretreatment (Yan et al., 2016) and steam explosion (Liu et al., 2014) are two other examples of physical/physicochemical pretreatment methodologies.

Even though the concentrated acid can almost completely break down cellulose at a lower temperature, the process is not practical as it produces significant waste that is harmful to the environment. Hence, the dilute acid pretreatment is employed as a favorable approach over other pretreatment methods due to its low cost, high efficiency in hydrolyzing hemicellulose into monomeric components, and generating structural modifications for improved enzyme accessibility and cellulose conversion (Loow et al., 2016). Sulfuric acid (H_2SO_4) is the most frequently used acid for dilute acid pretreatment, while nitric acid (HNO_3), hydrochloric acid (HCl), or phosphoric acid (H_3PO_4) can be used as well (Xu and Huang, 2014). However, this process also has several disadvantages, such as the need for expensive corrosion-resistant equipment or the neutralization of acidic hydrolyzates before the fermentation of sugars (Zheng et al., 2009).

Alkaline pretreatment can eliminate the need for costly materials and specialized designs for corrosion resistance or strong reaction conditions; it is performed at moderate conditions, sometimes even at room temperature, by soaking the material in a sodium hydroxide (NaOH) or ammonium hydroxide (NH_4OH) solutions. Some alkaline pretreatment techniques can also allow the recovery and reuse of chemical reagents (Kim et al., 2016). However, the efficiency of alkaline pretreatment depends on the substrate; generally, it is more successful on hardwood, herbaceous crops, and agricultural leftovers with low lignin content (Zheng et al., 2009). The primary downside of this technique is the formation of considerable amounts of salts, which limit microbial growth and ethanol fermentation in the next stages if they are not effectively removed (Liu et al., 2019). Ionic liquids (IL, high amount of organic cation with a small amount of inorganic anion), deep eutectic solvents (DES, mixtures of Lewis and Bronsted acids and bases), organosolv (organic solvents, *e.g.*, ethanol, methanol, butanol, acetone) methods can also be used (Galbe and Wallberg, 2019).

Biological pretreatment is also an environmentally friendly and economically promising alternative; in this method, the microorganisms such as brown, white, and soft rot fungi can degrade lignin and hemicelluloses from LCB (Sindhu et al., 2016). The main advantage of this process is that there is no requirement for chemical recycling, and no harmful substances are released into the environment. However, it has disadvantages like the need for very long reaction times due to slow degradation rate and the loss of significant amounts of biomass during the process (Liu et al., 2019). Although the overall aim of the pretreatment process, which is to maximize the release of fermentable sugars while limiting the inhibitor formation, is common, the best way to achieve this is highly dependent on the type of biomass (*i.e.*, chemical and physical properties of the biomass) (Ravindran and Jaiswal, 2016; Vollmer et al., 2022).

2.2. Hydrolysis

Following the pretreatment phase, the hydrolysis process occurs, which can be classified as acid and enzymatic hydrolysis (Lugani et al., 2020). Enzymatic hydrolysis has a lower environmental impact and inhibitor formation, while acid hydrolysis is faster (Vani et al., 2015). In enzymatic hydrolysis, because LCB is composed of cellulose, hemicellulose, and lignin, a cocktail of enzymes containing cellulase (*i.e.*, cellobiohydrolases, endo-glucanases, b-glucosidases), hemicellulase (*i.e.*, endo-xylanases, b-xylosidases, xyloglucanases), and lignin-degrading enzymes is required (Agrawal et al., 2021). Since the process uses soluble enzymes to break down insoluble substrates, it is a heterogeneous reaction system that is influenced by a variety of parameters such as lignin and hemicellulose content, cellulose crystallinity, degree of polymerization, accessible surface area, and pore volume (Zhao et al., 2021).

During the hydrolysis process, the cellulose is broken down into glucose while hemicellulose is separated into 5-carbon sugars (*i.e.*, arabinose and xylose); however, acetic acid is also produced as a byproduct of the hydrolysis of hemicellulose limiting the microbial development and ethanol production; this can be considered as the major disadvantage of hydrolysis process (Scheller and Ulvskov, 2010). Another disadvantage is the high energy consumption of the process because of the lignin present in the reaction, which consumes reactor space and creates a need for extra mixing to homogeneously suspend the fermentation broth during the enzymatic hydrolysis and fermentation stages (Liu et al., 2019).

2.3. Fermentation

A soup of hexose and pentose sugars is produced at the end of the hydrolysis process. The conversion of glucose to ethanol is simple and uncomplicated, but the others are not. Various microbial populations are needed for the fermentation of different sugars; however, each has different optimum growth conditions (Kucharska et al., 2018). Microbes that naturally ferment all these sugars also have a low tolerance for bioreactor

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

conditions due to toxin buildup. In addition, during the process of sugar fermentation, microorganisms tend to utilize one type of sugar (usually glucose) over others (Kim et al., 2010). For example, *Saccharomyces cerevisiae* is one of the most commonly used microorganisms in fermentation that cannot naturally utilize xylose (Jahanbakhshi and Salehi, 2019). Even though those microorganisms may utilize pentose sugars, the glucose generated from cellulose often inhibits the catabolism of these sugars (Zhao et al., 2021). All these make it difficult to develop and control the fermentation of LCB as a feedstock; the incomplete conversions and slow enzyme reactions also complicate the process and reduce the ethanol yield.

The efficiency of pretreatment, hydrolysis, and fermentation processes, together with LCB characteristics, are important for producing bioethanol at competitive prices (Qiao et al., 2022). High efficiency, low cost, and low level of inhibitory byproducts using greener pretreatment solutions are among the current research focuses (Sidana and Yadav, 2022). Fermentative microorganisms are also needed in the field, requiring more research for discovering efficient and robust microbial consortiums or constructing genetically engineered strains (Culaba et al., 2022). The selection of raw material is also another important factor affecting the cost of the bioethanol production process, as the composition has a direct effect on pretreatment cost and fermentable sugar content as well as the abundance of the biomass in the region of consideration (Smuga-Kogut et al., 2021). By taking all factors into account, it becomes hard to find the optimal solution for bioethanol production for different regions of the world. Traditional mathematical models can solve optimization problems with first-order equations (Sousa Jr et al., 2011). However, the variability is always high in biological systems; hence the generalization capacities of these models are not always sufficient. On the other hand, more generalizable solutions can be developed with the use of ML algorithms, which can overcome the nonlinearities and high level of complexity of the biological processes, especially if more research and experimental effort is dedicated to producing high-quality, reproducible data (Wang et al., 2022).

3. Bibliometric evaluation of lignocellulosic biofuel area

Bibliometric evaluation of scientific literature is widely performed in different areas of science and becoming a common research tool in specified research fields to connect relations among various concepts and research disciplines as well as to discover global research trends (Yaoyang and Boeing, 2013). To understand the trends in the "lignocellulosic biofuel" research field, bibliometric evaluation was done by analyzing the "author keywords" of the articles in the literature. For this purpose, the Web of Science database was used with the search term *lignocellulosic biofuel*, and a bibliometric study was carried out with a total of 6853 publications.

Research interest in the field is assessed by the number of articles published yearly. It was found that articles related to lignocellulosic biofuel are increasing year by year as expected (**Fig. 4a**). However, this trend is common in different research fields as the total number of SCI-indexed publications is also increasing (Yaoyang and Boeing, 2013). To discover the assistance of ML in the field, another search was conducted with the term; *lignocellulosic machine learning* and compared with *lignocellulosic biofuel*, as shown in **Figure 4a**. ML inclusion in the field was observed to be increasing in number and expanding its relative share in total publications.

"Author keywords" are extracted from the publications and categorized concerning the type of biofuel, feedstock, and conversion method to understand the trends in the area. A data cleaning step was conducted by combining the duplicated and synonymous keywords. Also, four-year moving averages were analyzed to eliminate any fluctuations in years. The result of categorized keyword distribution for four-year averages is presented in **Figures 4b-d** to uncover shifts in the research trend in the field.

Second-generation biofuels can be produced using various conversion methods, including hydrolysis-fermentation, pyrolysis, hydrothermal conversions, and other biological processes. Biofuels such as biogas, biohydrogen, bioethanol, biomethanol, and biodiesel can be produced using these conversion processes (Kucharska et al., 2018). To understand the trends in the lignocellulosic-based biofuel type, keywords are categorized with respect to the main biofuel categories: *bioethanol*, *biogas/biohydrogen*, *biodiesel*, *biobutanol*, *biochar*, and *fermentable sugar*. As shown in **Figure 4b**, *bioethanol* is the most studied lignocellulosic biofuel in each period. It is also observable that *biochar* and *biogas* are gaining more attention, as almost half of the related papers in these fields have been published in the last 4 years. As shown in **Figure 4b**, the trend of the conversion methods also agrees with the product-related keywords. *Pretreatment*, *hydrolysis*, and *fermentation* are the most frequently used keywords, as they constitute the main pathway for bioethanol production. However, although fermentation is more studied in total than pyrolysis, this gap is getting closer each year. Also, each year, *anaerobic digestion*, *hydrothermal liquefaction*, and *hydrothermal carbonization* increase their individual share in keywords.

The type of lignocellulosic feedstock utilized for biofuel production is critical for efficient and economic conversion. LCB can be categorized as agricultural and forest residues, forestry products, dedicated energy crops, municipal solids, and industrial waste (Qiao et al., 2022). The most commonly studied feedstocks in literature are categorized and shown in **Figure 4c**. The *Agricultural residue* category has the highest focus in the research area. *Microalgae*, although it is not a lignocellulosic material in nature, has a strong presence also in lignocellulosic biofuel-related articles and gaining more attention in recent years. It is also found that the number of different feedstocks tested increases yearly.

Categories and keywords related to bioethanol are analyzed further, and it was found that the use of *ionic liquids* (IL) in pretreatment is the most frequently appearing keyword indicating its recent popularity (**Fig. 4d**). *Dilute-acid pretreatment* (DA) was the second most studied one between 2011-2018; however, in the last 4 years, *organosolv* (OV), *steam explosion* (SE), and *microwave* (MW) treatment received more attention than *dilute-acid pretreatment*. The increase in the research interest in *deep eutectic solvents* (DESs) is also worth mentioning, as their frequency has almost doubled in the last four years. The most commonly studied microorganisms are also presented in **Figure 4d**; for hydrolysis, the research interest is focused mainly on *Trichoderma reesei*, *Clostridium thermocellum*, and *Aspergillus niger* while the interest in fermentation is more diverged even though *S. cerevisiae* is the choice for fermentation throughout the years.

4. Machine learning in lignocellulosic ethanol

As the review articles presented in **Table 1** indicate, ML has been used in various subfields of biofuel research in recent years. The optimization of process variables to produce biodiesel from algal oil (Kumar et al., 2018; Franco et al., 2019), estimation of biochar productivity and carbon content based on pyrolysis data from LCB (Zhu et al., 2019), optimization of process parameters for anaerobic fermentation of corn stalk (Dong and Chen, 2019), prediction of the physical and chemical properties of biodiesel based on its fatty acid content (Alviso et al., 2020), and feasibility of ML algorithms for estimating biodiesel purity (Moayedi et al., 2020) are some examples of such works in the literature. We have also published several ML applications on big databases extracted from the scientific literature, such as the evaluation of the catalytic activity of solid acid catalysts for the transesterification process (Alper Tapan et al., 2016), investigating the essential parameters of algal biomass and lipid production (Coşgun et al., 2021), analysis of biomass and lipid productivities of an oleaginous yeast namely *Yarrowia lipolytica* (Coşgun et al., 2022), and modeling biodiesel properties (*i.e.*, cetane number, cold filter plugging point and oxidation stability) over biodiesel samples (Suvarna et al., 2022). All those works indicate that, as far as ML is concerned, biofuel research is too diverse to analyze in a single communication; hence, in this review, we cover only the works directly related to lignocellulosic bioethanol production through the fermentation route.

For consistency, academic databases (*i.e.*, Web of Science, Scopus, and Google Scholar) were searched with keywords *lignocellulosic bioethanol* and *machine learning* (supplemented by keywords such as *data mining* and names of ML algorithms). After the preliminary and comprehensive examination, 43 articles were retrieved to represent the subject. It is also worth mentioning that this study is limited to ML studies focused on bioethanol production from LCB.

Figure 5 summarizes the articles presented in this work, with the numbers in the figure denoting the number of articles. **Figure 5a** shows the publication years of the articles, which shows an increasing trend in ML studies in the lignocellulosic bioethanol field, even though there are fluctuations due to the small data size. The distribution of data size used in these works is given in **Figure 5b**. It is indicated that the majority of the studies have data sizes ranging from 10 to 30 data points, likely due to the time-consuming nature of experimental work in the field. As a consequence

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Fig. 4. Publications trends in lignocellulosic biofuel through the years: **(a)** total and machine learning (ML)-related number of lignocellulosic biofuel articles; **(b)** keyword distribution of biofuel and conversion processes; **(c)** keyword distribution of biomass source; and **(d)** keyword distribution of pretreatment methods, and microorganisms used for fermentation and hydrolysis. Abbreviations: HTL: hydrothermal liquefaction; HTC: hydrothermal carbonization; SCB: sugarcane bagasse; LCW: lignocellulosic waste; IL: ionic liquids; STE: steam explosion; OV: organosolv; DES: deep eutectic solvents; MW: microwave; DA: dilute-acid; and US: ultrasound.

of the small data size, the number of descriptors is also small (*i.e.*, 2 to 5) in many works, as depicted in Figure 5c; those are the variables related to biomass characteristics, and operational conditions such as time, temperature and pH depending on the steps (*i.e.*, pretreatment, hydrolysis, and fermentation) involved and technology used. Figure 5d shows the choice of ML algorithm in the studies. The output variables are also given in Figure 5e; although most studies focus on direct outputs, such as bioethanol, fermentable sugar, and glucose, some studies concentrate on process efficiency-related outputs. Studies conducted throughout the fermentation process primarily focus on predicting and optimizing the input variables for bioethanol production, while fermentable sugar and glucose are the common output variables for the hydrolysis process.

The details of reviewed papers involving the pretreatment, hydrolysis, and fermentation steps are presented in Tables 2-4, respectively, while the major patterns observed in these papers are briefly discussed below with representative examples. Tables are categorized depending on the corresponding step of the input variables used in the ML modeling for a comprehensive understanding of the studies. Articles with variables only from the pretreatment step are summarized in Table 2. On the other hand, articles that include inputs from the hydrolysis step but exclude the fermentation step are summarized in Table 3. In Table 3, studies are categorized into; models that include variables from pretreatment and hydrolysis steps and models that only include hydrolysis step variables. In Table 4, articles that include the fermentation step inputs into the ML models are summarized with categorization performed in the same way as in Table 3.

For a comprehensive picture, the unitless performance metrics (i.e., R^2) of all ML models studied in the research articles mentioned in Tables 2-4 are summarized in Figure 6. The average performance of models is above the R^2 value of 0.90, which suggests the high predictive performance of the models. Although the number of models (represented as n) is not enough to make clear inferences, it can be concluded that the addition of pretreatment step inputs into models that are only built with hydrolysis step inputs increases the model performance. To understand the maximum achievable results of the major output variables (*i.e.*, bioethanol, fermentable sugar, and glucose concentration), the prediction results of ML-assisted modeling and optimization studies are given in Figure 7. The results are separated with respect to the LCB used in the models, as the bioethanol, fermentable

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

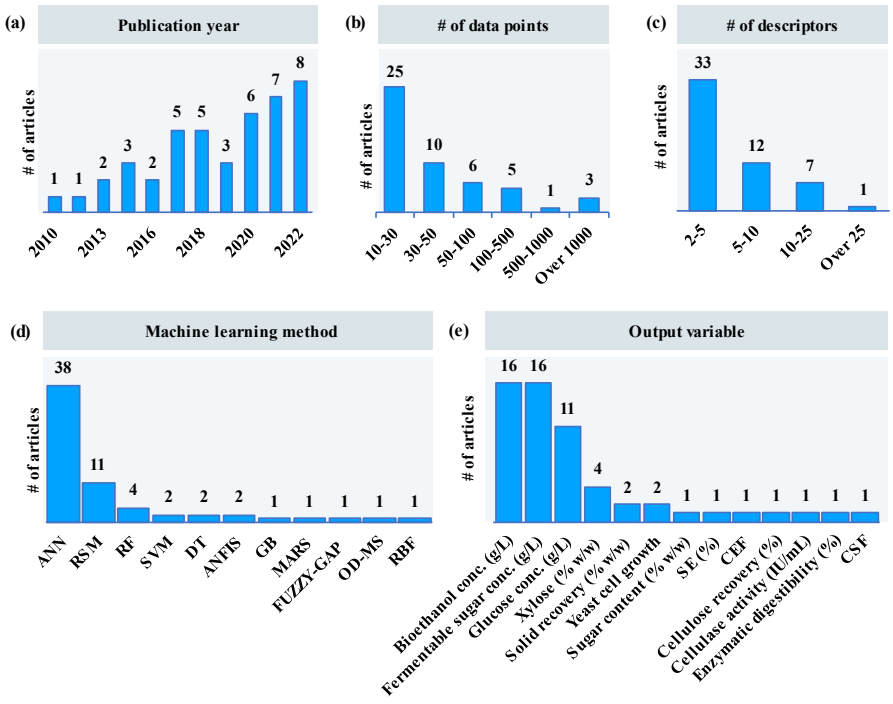


Fig. 5. ML trends for lignocellulosic ethanol production in papers covered in this review: **(a)** distribution over publication dates; distribution of **(b)** data sizes; **(c)** the number of descriptors; **(d)** ML algorithms used; and **(e)** output variables used in ML models. Abbreviations: ANN: artificial neural network; RSM: response surface methodology; RF: random forest; SVM: support vector machine; DT: decision tree; ANFIS: adaptive neuro-fuzzy inference system; GB: gradient boosting; MARS: multivariate adaptive regression splines; FUZZY-GAP: fuzzy system; OD-MS: optimized decision-making system; RBF: radial basis functions; SE: saccharification efficiency; CEF: cellulose enrichment factor; and CSF: combined severity factor.

Table 2.
Summary of studies in which machine learning was used for lignocellulosic bioethanol production, involving the pretreatment step.

Lignocellulosic Biomass	Pretreatment	Output Variable [d]	Data Size	Descriptors (# of Descriptors)	ML Method [e]	Performance Measures [f]	Reference
45 types of biomass	Ionic liquids	CEF	520 (literature data)	(23) biomass characteristic, pretreatment condition, main ILS identity, co-ILS identity, catalyst loading	SVM	R^2= 0.8998, RMSE= 0.2808	Phromphithak et al. (2021)
					GB	R^2= 0.9169, RMSE= 0.2556	
					RF	R^2= 0.9363, RMSE= 0.2238	
		SR (% w/w)			SVM	R^2= 0.7953, RMSE= 0.0830	
					GB	R^2= 0.8200, RMSE= 0.0778	
					RF	R^2= 0.8246, RMSE= 0.0768	
Rice straw	Microwave-assisted alkali treatment; NaOH	Cellulose recovery (%)	20	(3) alkali conc., microwave irradiation time, and strength	ANN	R^2= 0.99998[a], 0.99998[b], 1[c], 0.99365	Parkhey et al. (2020)
Water hyacinth	Dilute-acid pretreatment; H_2SO_4	Xylose (mg/g)	30	(4) temperature, acid concentration, treatment time, residence time	ANN	R^2= 0.9995	Das et al. (2016)
Mixed vegetable waste biomass	Dilute-acid pretreatment	Fermentable sugar conc. (mg/g)	29	(3) reaction time, reaction temperature, acid conc.	ANN	R^2= 0.921-0.986, RMSE= 0.0123-0.0166	Dharmalingam et al. (2022)
			58		RSM	R^2= 0.7686-0.9368	
Pennisetum grass	Alkali pretreatment; NaOH	Fermentable sugar conc. (mg/g)	16	(3) pretreatment temperature, acid conc., soaking time	ANN	R^2= 0.98, RMSE= 0.027	Mohaptra et al. (2016)
		Cellulose (mg/g)				R^2= 0.96, RMSE= 0.08	
		Hemicellulose (mg/g)				R^2= 0.89, RMSE= 0.205	
		Acid-soluble lignin (%)				R^2= 0.92, RMSE= 0.33	
3 types of oil palm	-	Lignin extraction (%)	15	(4) temperature, time, particle size range, solid loading	RSM	R^2= 0.8805, RMSE= 4.784	Rashid et al. (2021)
					ANN	R^2= 0.9933, RMSE= 1.129	
Hardwood (*Leucaena leucocephala*)	Organosolv treatment; glycerol	Fermentable reducible sugar (g/g)	17	(3) catalyst conc., duration, temperature	RSM	R^2= 0.996, RMSE= 5.564	Singhal et al. (2018)
			17*3		ANN	R^2= 0.998, RMSE= 3.630	

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Table 2.
continued.

Lignocellulosic Biomass	Pretreatment	Output Variable [d]	Data Size	Descriptors (# of Descriptors)	ML Method [e]	Performance Measures [f]	Reference
Wheat straw	Dilute-acid pretreatment	Xylose yield (%)	17	(3) reaction temperature and time, acid conc.	GPR	R^2= 0.999	Vollmer et al. (2022)
Cassava peels	Thermal-assisted dilute-acid pretreatment; HCl	Fermentable sugar conc. (g/L)	49	(5) soaking temperature and time, autoclave duration, HCl conc., solid loading	ANN	R^2= 0.82	Aruwajoye et al. (2022)
					RF	R^2= 0.64	
					DTR	R^2= 0.99	
		CSF			ANN	R^2= 0.93	
					RF	R^2= 0.77	
					DTR	R^2= 0.68	
Olive tree biomass	Inorganic salt-based treatment; $FeCl_3$	SR, glucose conc. (g/L)	15	(3) pretreatment duration and temperature, $FeCl_3$ conc.	RSM	R^2= 0.77[a], RMSE= 2.41[a], 5.52[b]	Charte et al. (2017)
					ANN	R^2= 0.81[a], RMSE= 2.14[a], 4.64[b]	
					FUZZY-GAP	R^2= 0.04[a], RMSE= 5.19[a], 4.09[b]	
					ANN	R^2= 0.07[a], RMSE= 5.12[a], 4.82[b]	
					SVM	R^2= 0.09[a], RMSE= 5.33[a], 4.86[b]	
Different LCBs	Dilute acid-assisted wet torrefaction; H_2SO_4	Glucose conc. (g/L)	49 sets	(13)	ANN	R^2= 0.9958	Chen et al. (2022)
					MARS	R^2= 0.929	
Napiergrass	Steam explosion followed by alkali pretreatment; NaOH	Enzymatic digestibility (%)	27	(3) steam explosion (temperature, time, and particle size)	ANN	R^2= 0.988[a], 0.975[b]	Chang et al. (2011)
Sago palm bark	Microwave-assisted dilute-acid treatment; H_2SO_4	Glucose (% w/w)	17	(3) microwave power, exposure time, solid loading	ANN-PSO	R^2= 0.9939	Ethaib et al. (2016)
		Xylose (% w/w)				R^2= 0.9479	
Sugarcane bagasse	Dilute-acid pretreatment; H_2SO_4	Glucose prod. (g/L)	32	(3) H_2SO_4 conc., solid ratio, autoclave residence time	ANN	R^2= 0.9774[a], 0.7939[b]	Gitifar et al. (2013)
	combined dilute-acid–ozonolysis method; H_2SO_4		36	(5) H_2SO_4 conc., solid ratio, autoclave residence time, moisture content, fixed-bed reactor residence time (ozonolysis time)		R^2= 0.9924[a], 0.9722[b]	
Oil palm empty fruit bunches	Ultrasonic-assisted organosolv treatment; ethanol	Fermentable sugar yield (g/g)	20	(3) temperature, time, sonication power	ANN	R^2= 0.90843[a], 0.8264[c]	Lee et al. (2020)

[a] The performance measured for the training set.
[b] The performance measured for the testing set.
[c] The performance measured for the validation set.
[d] CEF: cellulose enrichment factor; SR: solid recovery; CSF: combined severity factor.
[e] SVM: support vector machine; GB: gradient boosting; RF: random forest; ANN: artificial neural network; RSM: response surface methodology; GPR: Gaussian process regression model; DTR: decision tree regression; FUZZY-GAP: fuzzy system; MARS: multivariate adaptive regression splines; ANN-PSO: artificial neural network (ANN) in combination with particle swarm optimization (PSO).
[f] RMSE: Root mean square error.

Table 3.
Summary of the studies in which machine learning was used for lignocellulosic bioethanol production, involving the hydrolysis step.

Lignocellulosic Biomass	Pretreatment	Hydrolysis	Output Variable [d]	Data Size	Descriptors (# of Descriptors)	ML Method [e]	Performance Measures [f]	Reference
Pretreatment & Hydrolysis								
Sugarcane leaf waste	Inorganic salt-based treatment; $ZnCl_2$	Enzymatic hydrolysis	Fermentable sugar yield (% w/w)	90	(8)	ANN	R^2= 0.97	Moodley et al. (2019)
Microalgal biomass	Alkali pretreatment; H_2O_2	Enzymatic hydrolysis	Carbohydrate conc. (g/g)	24	(3) wastewater conc., H_2O_2 and enzyme activity	ANN	R^2= 0.99972[a], 0.99435[c], 0.99937[b], 0.9791	Onay (2022)
						RSM	R^2= 0.91. MSE= 0.78	
Sugarcane bagasse	Dilute-acid pretreatment; H_2SO_4	Enzymatic hydrolysis	Glucose yield (%)	3049	(6) pretreatment time, initial biomass conc., acid conc., enzyme conc., hydrolysis time, substrate loading in hydrolysis	ANN	MSE= 6.8, R^2= 0.987	Plazas Tovar et al. (2018)

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Table 3.
continued.

Lignocellulosic Biomass	Pretreatment	Hydrolysis	Output Variable [d]	Data Size	Descriptors (# of Descriptors)	ML Method [e]	Performance Measures [f]	Reference
Lignocellulosic biomass	Dilute-acid pretreatment	Enzymatic hydrolysis	Fermentable sugar conc. (g/L)	107 (literature data)	(9) three major constituents of biomass composition, pretreatment conditions (acid conc., temperature, time), the ratio of cellulose to lignin, cellulase concentration, the severity of acid pretreatment	ANN PRM	R^2= 0.997[a], 0.984[c], 0.967[b] R^2= 0.963	Haldar et al. (2023)
Corn stover	Dilute-acid pretreatment; HCl, H_2SO_4, H_3PO_4	Enzymatic hydrolysis	Phenolic contents and glucose yield	77	(6) acid conc., pretreatment temperature, residence time, solid-to-liquid ratio, kinds of inorganic acids, enzyme loading dosage	ANN	R^2= 0.904 (phenolic conc.) and 0.906 (glucose conc.)	Luo et al. (2021)
			Hydrolysis					
Pumpkin peel waste	-	Enzymatic hydrolysis	Fermentable sugar conc. (g/L)	30	(4) hydrolysis time, substrate to liquid ratio, α-amylase conc., amyloglucosidase conc.	ANN RSM	R^2= 1[a], 0.99979[c], 0.99988[b] R^2= 0.988	Chouaibi et al. (2020)
Cocoa pod shell	Microwave	Acid hydrolysis	Fermentable sugar conc. (g/L)	12	(2) cocoa pod shell weight, H_2SO_4 conc.	RSM ANN	R^2= 0.89 R^2= 0.94	Shet et al. (2018a)
Non-edible seed cake	Autoclave	Acid hydrolysis	Fermentable sugar conc. (g/L)	12	(2) time, HCl conc.	ANN RSM	R^2= 0.975, RMSE= 1.078 R^2= 0.888, RMSE= 2.139	Shet et al. (2018b)
Waste broken rice	-	Enzymatic hydrolysis	Fermentable sugar yield (g/g)	30	(4) temperature, time, pH, and enzyme dosages	ANN RSM	R^2= 0.993, RMSE= 0.078 R^2= 0.987, RMSE= 0.102	Mondal et al. (2021)
Peanut shell	Combination of different alkali, dilute acid, steam explosion, and alkali steam-assisted sequential acid techniques	Enzymatic hydrolysis	Fermentable sugar conc. (g/L)	45	(3) temperature, substrate conc., and spore dosage	ANN	R^2= 0.929	Ganguly and Das (2022)
Rice straw	Microwave-assisted alkali treatment; NaOH	Enzymatic hydrolysis	SE (%)	30	(4) substrate conc., the enzyme load, temperature, and Tween-80 conc.	ANN	R^2= 0.99191[a], 0.92605[b], 0.98104[c], 0.947	Parkhey et al. (2020)
Rice straw	Alkali pretreatment; NaOH	Enzymatic hydrolysis	glucose and xylose yield (g/L)	120	(2) biomass loading and particle size	ANN	R^2= 0.99[a], 0.98[c], 0.97[b]. MSE= 0.567[a], 0.949[c], 1.555[b]	Vani et al. (2015)
Apple pomace	-	Enzymatic hydrolysis	Glucose and fermentable sugar conc. (g/L)	81	(4) substrate loading, enzyme loading, temperature, initial pH	ANN	R^2= 0.99	Gama et al. (2017)
Corn bran, wheat bran, and pine sawdust	-	Acid hydrolysis	Glucose conc. (g/L) Fermentable sugars conc. (g/L)	70	(4) hydrolysis temperature, H_2SO_4 conc., acid solution/feedstock ratio, hydrolysis time	RBF-PSO	R^2= 1.000, 1.000, and 0.995 for wheat bran, corn bran, and pine sawdust R^2= 0.979, 0.859, and 0.992 for wheat bran, corn bran, and pine sawdust	Giordano et al. (2013)
Sweet sorghum	-	Enzymatic hydrolysis	Fermentble sugar conc. (g/L)	29	(4) substrate loading, α-amylase conc., amyloglucosidase conc., stroke speed	ANN	R^2= 0.994	Sebayang et al. (2017)
Sugarcane bagasse	Alkali pretreatment; H_2O_2	Enzymatic hydrolysis	Glucose yield (% theoretical max.)	480	(3) cellulase, β-glucosidase, time	ANN	Validation gave acceptable performance measures	Rivera et al. (2010)

[a] The performance measured for the training set.
[b] The performance measured for the testing set.
[c] The performance measured for the validation set.
[d] SE: saccharification efficiency.
[e] ANN: artificial neural network; RSM: response surface methodology; PRM: polynomial regression model; RBF-PSO: radial basis functions (RBF) in combination with particle swarm optimization (PSO).
[f] RMSE: Root mean square error.

sugar, and glucose concentrations are highly dependent on the nature of the feedstock.

As the initial phase in ethanol production from LCB, several researchers concentrate on enhancing the pretreatment procedure. The output variables are generally cellulose recovery, while the descriptors are biomass characteristics and pretreatment conditions in these works. For instance, Phromphithak et al. (2021) modeled cellulose enrichment factor (CEF) and solid recovery (SR) by support vector machine (SVM), gradient boosting (GB), and random forest (RF) using 45 types of biomass and 80 kinds of solvents with 520 data entries gathered from the literature. It was shown that RF has higher predictive performance for CEF and SR (% w/w), while the other ML algorithms performed better for CEF. Similarly, the effect of

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Table 4.
Summary of the studies in which machine learning was used for lignocellulosic bioethanol production, involving the fermentation step.

Lignocellulosic Biomass	Pretreatment	Hydrolysis	Fermentation	Output Variable	Data Size	Descriptors (# of Descriptors)	ML Method [d]	Performance Measures [e]	Reference
Pretreatment, Hydrolysis & Fermentation									
Buckwheat straw and biomass from wastelands	Ionic liquids	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	144	(14) biomass composition, type and amount of ionic liquids, types of enzymatic preparations, glucose content	ANN	R^2= 0.93[a], 0.78[c]	Smuga-Kogut et al. (2021)
							RF	R^2= 0.93[a], 0.94[c]	
						(11) biomass composition, types and amount of ionic liquids, time, combinations of two enzymes, glucose content	ANN	R^2= 0.99[a], 0.88[c]	
							RF	R^2= 0.93[a], 0.96[c]	
Waste potato mass	Ultrasonication	Acid hydrolysis	*S. Cerevisiae*	Bioethanol yield (g/L)	17	(3) HCl conc., ultrasonication time, S. cerevisiae conc.	RSM	RMSE= 0.201, R^2= 0.9628	Suresh et al. (2020)
							ANN	RMSE= 0.106, R^2= 0.979	
		Enzymatic hydrolysis			17	(3) α-amylase conc., ultrasonication time, S. cerevisiae conc.	RSM	RMSE= 0.235, R^2= 0.9513	
							ANN	RMSE= 0.124, R^2= 0.9587	
Hydrolysis & Fermentation									
Forest products and agricultural residues	-	Enzymatic hydrolysis	*S. Cerevisiae*	Glucose yield (g/L) Bioethanol yield (g/L)	300 datasets	(11) biomass composition; saccharification time, temperature, pH, shaking speed; fermentation time, temperature, pH, shaking speed	OD-MS	Overall accuracy = 95%	Vinitha et al. (2022)
Sugarcane bagasse	-	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	17 runs, 1560 data	(5) temperature, enzyme conc., biomass load, inoculum size, and time	ANN	R^2= 0.92[a], 0.90[b]. RMSE= 0.68[a], 0.78[b]	Fischer et al. (2017)
							RF	R^2= 0.92[a], 0.91[b]. RMSE= 0.77[a], 0.87[b]	
							DT	error = 12.2%	
Fermentation									
Marine macroalgae	-	Acid hydrolysis	*S. Cerevisiae*	Bioethanol prod. (g/g RS)	Around 80	(6) substrate conc., fermentation time, inoculum size, temperature, agitation speed, pH	ANN	R^2= 0.94[a], 0.99[b], and 0.99[c]. MSE= 0.00735	Dave et al. (2021)
Intermediates and byproducts of sugar beet processing	*not necessary, not lignocellulosic	-	*S. Cerevisiae*	Bioethanol content (% v/v) Yeast cell number (10^8 cells ml/L) Sugar content (% w/w)		(3) fermentation time, starting sugar content, substrate type	ANN	R^2 from 0.823 to 0.999 R^2 from 0.692 to 0.993 R^2 from 0.929 to 0.999	Grahovac et al. (2016)
Pumpkin peel wastes	-	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	30	(4) growth temperature, pH, agitation speed, yeast conc.	ANN	RMSE= 0.7968[a], 0.05924[c] and 0.989[b]. R^2= 0.984306[a], 0.9986[c], 0.99247[b]	Chouaibi et al. (2020)
							RSM	R^2= 0.9762	
Corn cobs and corn stovers hydrolysate	-	Acid or enzymatic hydrolysis	*S. Cerevisiae*	Cell growth and ethanol prod.	48	(208) volatile components GC-MS peak data	ANN	Learning and validation losses, 0.033 and 0.507	Konishi (2020)
Sugarcane	-			Bioethanol conc.	3400 data (200 days)	(7) related to the different areas of the fermentation unit, from the composition of the must to the centrifugation of the wine	ANN	R^2= 0.91, MSE= 0.26	Pereira et al. (2020)
Corn steep liquor	-		*S. Cerevisiae*	Bioethanol prod. (g/L)	46	(5) substrate conc., pH, time, temperature, inoculum size	RSM	R^2= 0.34, RMSE= 1.29	Taiwo et al. (2018)
							ANN	R^2= 0.98, RMSE= 0.19	
			Instant dry yeast		46	(5) substrate conc., pH, time, temperature, inoculum size	RSM	R^2= 0.98, RMSE= 0.97	
							ANN	R^2= 0.99, RMSE= 0.29	

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Table 4.
continued.

Lignocellulosic Biomass	Pretreatment	Hydrolysis	Fermentation	Output Variable	Data Size	Descriptors (# of Descriptors)	ML Method [d]	Performance Measures [e]	Reference
Watermelon waste	-	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol prod. (% w/w)	27	(2) agitator speed, yeast amount	ANN ANFIS	R^2= 0.9895 R^2= 0.9993	Jahanbakhshi and Salehi (2019)
Microalgal biomass	Alkali pretreatment; H_2O_2	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	36	(5) substrate conc., inoculum, fermentation time, pH, and temperature	ANN RSM	R^2= 0.99971[a], 0.74285[c], 0.93635[b], 0.90114 R^2= 0.94	Onay (2022)
Manihot esculenta Crantz YTP1 stem	Dilute-acid pretreatment; CH_3COOH, HNO_3	Enzymatic hydrolysis	*Z. mobilis*	Cellulase activity (IU/mL) Bioethanol yield (g/L)	30	(4) pH, temperature, agitation, and time	ANN	R^2= all 0.990, MSE= 0.2654, RMSE= 0.5151 R^2= all 0.979, MSE= 0.4324, RMSE= 0.6575	Selvakumar et al. (2018)
Oil palm trunk sap	-		*S. Cerevisiae*	Bioethanol conc. (g/L)		(4) fermentation time, pH, temperature, total sugar	ANFIS	R^2= 1[a], 0.9991[b], 0.99975[c]	Ezzatzadegan et al. (2021)
Breadfruit starch hydrolysate	-		Instant dry yeast	Bioethanol yield (% v/v)	17	(3) reducing sugar conc., fermentation time, pH	ANN	R^2= 0.9995	Betiku and Taiwo (2015)
Sweet sorghum	-	Enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	17	(3) yeast conc., reaction temperature, agitation speed	ANN	R^2= 0.987	Sebayang et al. (2017)

[a] The performance measured for the training set.
[b] The performance measured for the testing set.
[c] The performance measured for the validation set.
[d] ANN: artificial neural network; RF: random forest; RSM: response surface methodology; OD-MS: optimized decision-making system; DT: decision tree; ANFIS: adaptive neuro-fuzzy inference system.
[e] RMSE: Root mean square error.

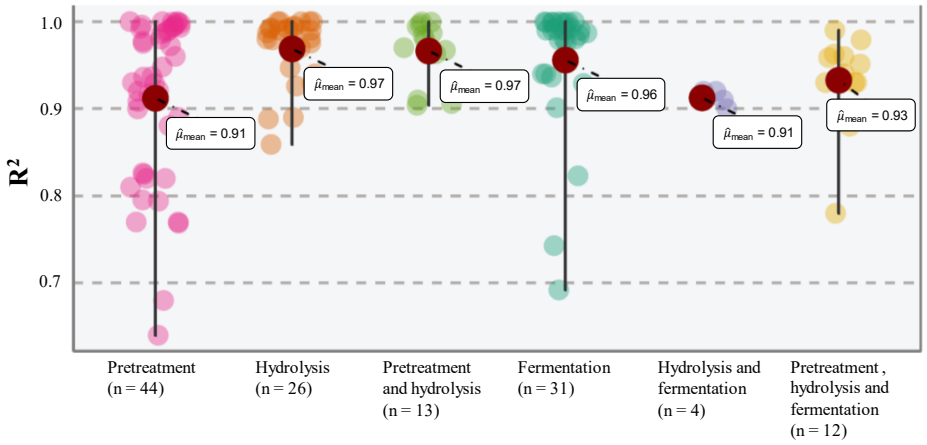


Fig. 6. R^2 values of ML models with respect to corresponding steps of the input variables used in the models (n: number of models, $\hat{\mu}_{mean}$: average of the R^2 values).

pretreatment conditions on cellulose recovery using rice straw was investigated by ANN models with the Levenberg-Marquadt back-propagation algorithm by Parkhey et al. (2020).

The cellulose content of LCB is transformed into fermentable sugars by hydrolysis. Among the studies focused on the conversion efficiency of LCB to fermentable sugars by ML models, some studied pretreatment and hydrolysis steps together. For example, Aruwajoye et al. (2022) studied both fermentable sugar concentration and combined severity factor (CSF), which represents the efficiency of the pretreatment method, using ANN, RF, and decision tree regression (DTR). They used soaking temperature, soaking time, autoclave duration, HCl concentration, and solid loading as descriptor variables and constructed models using 49 experimental data. It was found that the most successful ML method varied depending on the output variable studied.

Moodley et al. (2019) and Lee et al. (2020) studied the effect of experimental conditions on fermentable sugar content and found the optimum operation conditions as 30 min sonication treatment with 192.5 W in 48.2 ºC resulting in 356 mg/g biomass, while Onay (2022) offered ANN and RSM models for carbohydrate concentration; Chen et al. (2022), Gitifar et al. (2013), and Plazas Tovar et al. (2018), on the other hand, modeled the glucose concentration as the output variable. Likewise, Ethaib et al. (2016) studied both glucose and xylose as fermentable sugars, whereas Chang et al. (2011) modeled enzymatic digestibility (%) by using inputs from both pretreatment and enzymatic hydrolysis steps. Charte et al. (2017) analyzed the solid recovery and glucose content *via* various ML methods.

There are also studies focusing on the hydrolysis step alone, even though these works also consider different output variables (single or multiple).

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

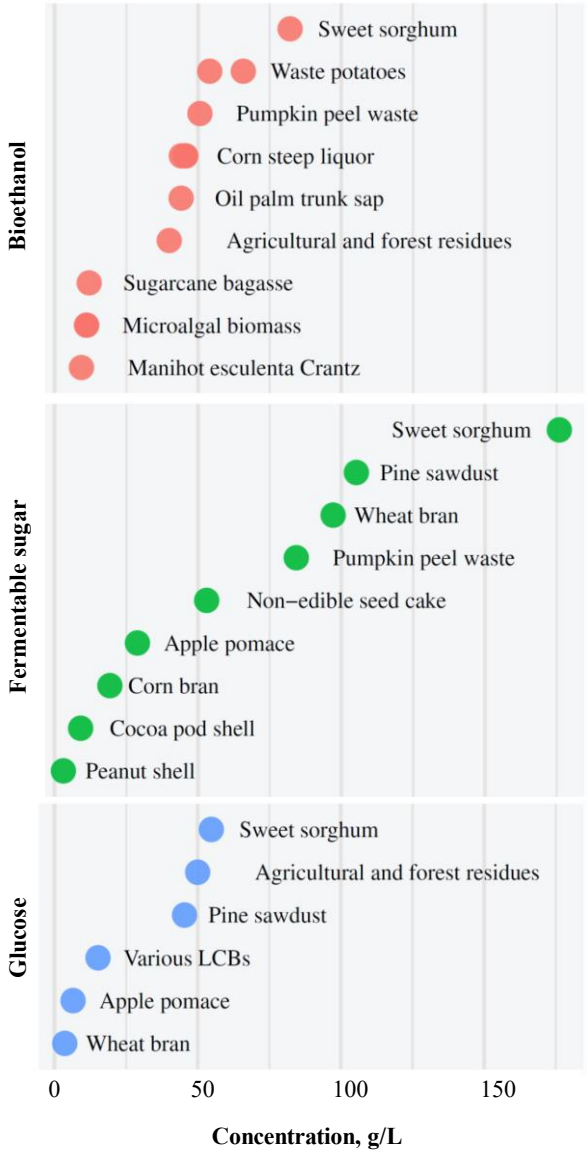


Fig. 7. Prediction results of maximum concentration of selected outputs through ML-assisted modeling and optimization with data from (Giordano et al., 2013; Fischer et al., 2017; Gama et al., 2017; Sebayang et al., 2017; Selvakumar et al., 2018; Shet et al., 2018a and b; Taiwo et al., 2018; Chouaibi et al., 2020; Suresh et al., 2020; Ezzatzadegan et al., 2021; Chen et al., 2022; Ganguly and Das, 2022; Onay, 2022; Vinitha et al., 2022). Abbreviation: LCB: lignocellulosic biomass.

For example, Rivera et al. (2010) studied glucose yield, while Vani et al. (2015) studied both glucose and xylose yield; both used ANN as the ML algorithms. Gama et al. (2017) and Giordano et al. (2013) also modeled glucose and fermentable sugar concentrations together. On the other hand, various researchers modeled fermentable sugar concentration and determined its maximum as 3.10 g/L (Ganguly and Das, 2022), 84.27 g/L (Chouaibi et al., 2020), 9.10 g/L (Shet et al., 2018b), 53.03 g/L (Shet et al., 2018a), 0.704 g/g (Mondal et al., 2021), and 175.94 g/L (Sebayang et al., 2017). As an additional example, Parkhey et al. (2020) studied the effects of enzymatic hydrolysis process variables on saccharification efficiency (SE) using ANN.

Fermentation, as the final phase of the three-step bioethanol production process from LCB, has been modeled by itself or together with the previous steps by various researchers. For example, Smuga-Kogut et al. (2021) and Suresh et al. (2020) modeled bioethanol production by using the variables belonging to three consecutive steps as the descriptors. Suresh et al. (2020) studied both acid and enzymatic hydrolysis with ANN using HCl and α-amylase concentration, ultrasonication time, and *S. cerevisiae* concentration as input for bioethanol yield. They also compared the results with response surface methodology (RSM) and found that ANN was superior in modeling bioethanol production. Smuga-Kogut et al. (2021) used three different ML methods: ANN, RF, and a hybrid model of both, and reported that the hybrid model provided more accurate results. The effect of hydrolysis and fermentation variables on bioethanol yield was also studied by Fischer et al. (2017) and Vinitha et al. (2022).

Many groups modeled the fermentation step only with the output variable of bioethanol concentration (all used different forms of ANN). On the other hand, Konishi (2020) studied cell growth and bioethanol production together, while Selvakumar et al. (2018) investigated cellulase activity and bioethanol yield; Grahovac et al. (2016), on the other hand, analyzed the bioethanol content, yeast cell number, and sugar content in the fermentation process.

5. Limitations and practical implications of the current work

Although we attempted to cover a sufficient number of papers involving a variety of aspects to provide an accurate representation of the current status of the pretreatment-hydrolysis-fermentation route for lignocellulosic bioethanol production, limitations and weaknesses are inevitable in such a review. First, we might have missed some significant works, as covering all related studies in a single review is impossible. Our restrictions on the scope and focus on bioethanol (no other product) from lignocellulose (no other raw material) *via* fermentation (no other processes) was necessary to see field-specific trends and make the review in manageable size; however, there is an obvious trade-off in this approach that we may miss the big picture and overlook some trends in biofuel production in general.

We think there are also some limitations and weaknesses arising from the current ML practice. One of the primary issues in the subject is the lack of data; unfortunately, sufficiently large datasets with high-quality data are rarely available. ML relies on statistical inference, requiring large datasets with reasonable accuracy. The researchers in the field either use their own experimental datasets, which are usually limited in size for reliable conclusions, or extract data from the literature, which contain significant levels of noise due to the non-standard nature of experimental conditions. In either case, the knowledge that can be extracted using ML is inevitably limited. There are also some common mistakes in ML applications that may lead to deficient and erroneous conclusions. For example, the ML algorithm is not always chosen by considering the knowledge to be extracted or the dataset's structure. Instead, it may be selected because of recent popularity providing only limited benefit if an unsuitable algorithm is selected. This may also be true for some of the articles we analyzed because it is not always simple to identify such issues unless researchers test alternative methods and describe them in their papers.

Another potential issue is that the models may be too large for the size of the data since the signs of overfitting are not always apparent, as in the case of simple regression. This typically occurs and goes unnoticed if an effective validation procedure is not implemented or the details of the procedure are not discussed in the paper. Even with the appropriate dataset size and effective algorithms and validation procedures, it is necessary to test a broad range of model hyperparameters to determine the optimal model structure that accurately represents the data. Occasionally, only a few sets of model hyperparameters are examined, particularly if the initial trials yield a satisfactory level of fitness.

Nevertheless, the limitations and weaknesses listed above are shared by all review papers of this type, and our paper will still make an important contribution to the field. We think that our review has four major implications in practice. First, it describes the current status, the patterns, and major research findings in the field through bibliometric analysis of the literature. Second, it provides consolidated results of representative works in literature for the readers to deduce their own conclusions. Third, as connected to the first two, our work may help to plan future experimental works by providing insight into the effects of descriptors, such that the focused nature of our work (bioethanol from lignocellulosic material *via* fermentation route) should help to identify some practical leverage points to improve the relevant processes further. Finally, the present work provides representative examples of ML applications for those who wish to

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

perform similar works. One of the most critical tasks in ML applications is the selection of descriptors; inspecting the descriptors lists and relative significances determined in various works will help the investigators identify the potential descriptors they should use. Additionally, the examples reviewed in this paper also direct the researchers to the data sources and speed up the execution of similar ML analyses.

6. Challenges and future perspectives

As also stated in the previous section as one of the major limitations in current works, the availability of a sufficiently large number of accurate data is one of the biggest challenges for ML applications in bioethanol production, and this will likely be the case in the near future as well.

ML requires a dataset that describes the physical process well. First, data should contain the desired information (*i.e.*, descriptors like physical and chemical properties of material or operational conditions should explain some critical performance measures). Second, the dataset should be sufficiently large and accurate for statistically reliable inferences. Construction of a sufficiently large and accurate dataset is one of the biggest challenges for ML applications in many fields; this also seems to be the case for lignocellulosic ethanol production, and as we stated in the previous section, it is also one of the major limitations of current applications. In fact, this may be more problematic for complex systems, such as lignocellulosic ethanol production, because a larger number of descriptors is required to represent such systems adequately, necessitating larger datasets for statistically reliable models. Another reason for this challenge in bioethanol production is that various alternative routes (like thermochemical or fermentation routes) or different configurations of the processes in the same route (like sequential or simultaneous hydrolysis and fermentation steps) are considered for lignocellulosic ethanol production, and none of them is regarded as the dominant route. Since the descriptors (sometimes performance measures) differ for dissimilar routes, the data from different process combinations differ. Hence, the availability of diverse processes and process configurations divides the efforts among the alternative routes and prevents the accumulation of sufficient data in any of them.

Furthermore, new material or process steps tested the first time create unique variables not reported by other papers. All these create significant difficulties for implementing ML, which relies on learning from existing relations in the data set; single or few data points having variables not shared by the others have limited use in ML analysis. The data seems to be a bigger problem for more complex configurations like performing hydrolysis and fermentation simultaneously because more descriptors will be needed to represent the combined process, which will require more data entries as well.

Another challenge seems to be the non-standard nature of cellulosic raw materials resulting in different products and yields, especially in the pretreatment and hydrolysis steps (Raj et al., 2022). Normally this should not be a problem for ML if all descriptors are clearly identified, and a sufficiently large number of data is available to smooth out the variations; however, in this field, datasets are typically small, and there is a substantial level of uncertainty (or at least variation) associated with the descriptors.

On the other hand, there are also efforts to overcome these challenges, and more can be expected in the future. One of these efforts is an approach called *transfer learning*, aiming to utilize the ML models and analysis developed for some fields to understand other similar fields (Kaya and Hajimirza, 2019). Although these are not easy to implement in practice, they may be beneficial for lignocellulosic ethanol production as well; for example, experiences and models developed for the fermentation of sugar from corn, which is a more established field, should provide some insights for the ML analysis of the fermentation step in lignocellulosic ethanol production even though some impurities (including inhibitors) exist for the cases related to lignocellulosic ethanol.

Using computational tools, especially density functional theory (DFT), is another option to create a dataset for ML analysis, and it is commonly employed in material research. The standard nature of the data created this way eases data sharing among the researchers; indeed, numerous databases like Material Project (Jain et al., 2016), OQMD (Kirklin et al., 2015), AFLOWLIB (Curtarolo et al., 2012), and Computational Material Repository (Landis et al., 2012) were constructed for this purpose. However, these tools and databases are mostly used for crystals and simple molecules; the current computational state may not be sufficient to generate the large number of data entries required for a process like fermentation. The use of experimental databases like Inorganic Crystal Structure Database (ICSD) (Bergerhoff et al., 1983), Pearson Crystal Data (Villars and Cenzual, 2007), Cambridge Structural Database (Allen, 2002), Crystal Open Database (Gražulis et al., 2009) or creation of a database for lignocellulosic ethanol production does not seem to be practical either. However, some sort of data-sharing mechanisms can still be implemented to improve the benefit of ML because larger datasets with more features always provide more detailed and accurate information in ML analysis. One way to do this is to develop some standard testing and reporting protocols, with the collaboration of researchers in the field, so data from various experimental works can be combined to create a sufficiently large amount of relatively uniform data. In the long run, computational tools like DFT can also be utilized in this field to understand the process and generate data considering the astonishing speed of progress in computational tools and algorithms.

Another approach that can be used for small datasets is reducing the number of descriptors (dimensionality reduction) because a lower number of descriptors requires smaller datasets; this can be done by eliminating insignificant descriptors (feature selection) or combining them into a smaller new descriptor set (feature extraction) (Alpaydin, 2020). Meanwhile, new ML algorithms and approaches for small datasets have also been investigated in recent years (Zhang and Ling, 2018; Feng et al., 2019; Ma et al., 2020). This trend will likely grow in the future and contribute to the research in lignocellulosic biofuels as well.

Finally, a concept called explainable ML has been discussed in recent years against the black box nature of ML models as one of the main weaknesses (and criticism) of the current ML applications (Suvarna et al., 2022). Although this concept is also hard to implement (like transfer learning), it is quite appealing because it aims to explain the reasons behind the results obtained by ML models. This approach may be more beneficial for complex systems like lignocellulosic ethanol because it helps to understand the relations among the descriptors and their impact on the outcome and allow to reduce their number (*e.g.*, reduction in the size of the dataset) by eliminating the insignificant descriptors, and make the use of small datasets easier.

7. Conclusions

Although LCB is the most abundant biomass source, converting it to ethanol is not an easy process and involves many sophisticated steps because of the nature of the LCB. In this article, first, the lignocellulosic bioethanol process was reviewed from several different angles, including the present state of research, underlying mechanisms, challenges, and obstacles. It was revealed that the pretreatment procedure is one of the most expensive steps with numerous approaches, including physical/physicochemical, acid/alkaline, solvent, and biological treatments. During the hydrolysis (which follows the pretreatment process), a cocktail of enzymes containing cellulase, hemicellulase, and lignin-degrading enzymes is necessary to break down the cellulose, hemicellulose, and lignin in the LCB. The hydrolysis process results in a soup of hexose and pentose sugars. The conversion of glucose (the main hexose sugar) to ethanol is straightforward, while the others are challenging.

In the second part of this work, a bibliometric analysis was performed to extract the trends of research interest in the field. It was found from this analysis that the inclusion of ML in the field is not only increasing but also expanding its relative share. Bioethanol was discovered to be the most researched lignocellulosic biofuel, while biochar and biogas have received increased attention in recent years, with nearly half of those studies published in the last four years.

Then, the implementation of ML approaches to assist in choosing the most suitable experimental conditions leading to the highest conversion *via* the most practicable route was reviewed in depth. It was observed that ANNs are the most commonly used algorithms (appeared in almost 90% of articles), followed by RSM (in about 25% of articles) and RF (in about 10% of articles). These numbers also indicate that most of the works in these articles are performed for the *prediction* task. *Bioethanol concentration* is the most common output variable to predict in fermentation steps, while *fermentable sugar* and *glucose concentration* are the most common output variables in hydrolysis. No such generalization was possible for pretreatment methods due to the diversity of the goals and the pretreatment process. The size of the datasets used in the analysis is usually small, while

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

the fitnesses of the models developed are usually high considering the R^2 values reported in the papers.

In addition, major challenges related to ML approaches were discussed in detail under three main steps: constructing the dataset, selecting and implementing ML algorithms, and interpreting the results. It was then concluded that due to the complexity and multi-step nature of the lignocellulosic ethanol production process, the availability of a sufficient amount of data would likely be a problem in the future. One way to improve data availability is by using standardized testing and reporting protocols within the field so that more data can be combined and used for ML analysis. New developments in ML, such as transfer learning, explainable ML, and algorithms allowing to work in small datasets, may also contribute to the development of the field.

References

[1] Agrawal, R., Verma, A., Singhania, R.R., Varjani, S., Di Dong, C., Kumar Patel, A., 2021. Current understanding of the inhibition factors and their mechanism of action for the lignocellulosic biomass hydrolysis. Bioresour. Technol. 332, 125042.

[2] Aguilar-Reynosa, A., Romaní, A., Ma. Rodríguez-Jasso, R., Aguilar, C.N., Garrote, G., Ruiz, H.A., 2017. Microwave heating processing as alternative of pretreatment in second-generation biorefinery: an overview. Energy Convers. Manage. 136, 50-65.

[3] Allen, F.H., 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr., Sect. B: Struct. Sci. 58(3), 380-388.

[4] Alpaydin, E., 2020. Introduction to Machine Learning, 4 (ed). The MIT Press.

[5] Alper Tapan, N., Yıldırım, R., Erdem Günay, M., 2016. Analysis of past experimental data in literature to determine conditions for high performance in biodiesel production. Biofuels, Bioprod. Biorefin. 10(4), 422-434.

[6] Alviso, D., Artana, G., Duriez, T., 2020. Prediction of biodiesel physico-chemical properties from its fatty acid composition using genetic programming. Fuel. 264, 116844.

[7] Aruwajoye, G.S., Faloye, F.D., Kassim, A., Saha, A.K., Kana, E.G., 2022. Intelligent modelling of fermentable sugar concentration and combined severity factor (CSF) index from pretreated starch-based lignocellulosic biomass. Biomass Convers. Biorefin.

[8] Aui, A., Wang, Y., Mba-Wright, M., 2021. Evaluating the economic feasibility of cellulosic ethanol: a meta-analysis of techno-economic analysis studies. Renew. Sust. Energy Rev. 145, 111098.

[9] Bannor B, E., Acheampong, A.O., 2019. Deploying artificial neural networks for modeling energy demand: international evidence. Int. J. Energy Sect. Manage. 14(2), 285-315.

[10] Bergerhoff, G., Hundt, R., Sievers, R., Brown, I.D., 1983. The inorganic crystal structure data base. J. Chem. Inf. Comput. Sci. 23(2), 66-69.

[11] Betiku, E., Taiwo, A.E., 2015. Modeling and optimization of bioethanol production from breadfruit starch hydrolyzate vis-à-vis response surface methodology and artificial neural network. Renewable Energy. 74, 87-94.

[12] Brandt, A., Gräsvik, J., Hallett, J.P., Welton, T., 2013. Deconstruction of lignocellulosic biomass with ionic liquids. Green Chem. 15(3), 550-583.

[13] Chang, C.W., Yu, W.C., Chen, W.J., Chang, R.F., Kao, W.S., 2011. A study on the enzymatic hydrolysis of steam exploded napiergrass with alkaline treatment using artificial neural networks and regression analysis. J. Taiwan Inst. Chem. Eng. 42(6), 889-894.

[14] Charte, F., Romero, I., Pérez-Godoy, M.D., Rivera, A.J., Castro, E., 2017. Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass. Comput. Chem. Eng. 101, 23-30.

[15] Cheah, W.Y., Sankaran, R., Show, P.L., Tg. Ibrahim, T.N.B., Chew, K.W., Culaba, A., Chang, J.S., 2020. Pretreatment methods for lignocellulosic biofuels production: current advances, challenges and future prospects. Biofuel Res. J. 7(1), 1115-1127.

[16] Chen, W.H., Lo, H.J., Aniza, R., Lin, B.J., Park, Y.K., Kwon, E.E., Sheen, H.K., Grafilo, L.A.D.R., 2022. Forecast of glucose production from biomass wet torrefaction using statistical approach along with multivariate adaptive regression splines, neural network and decision tree. Appl. Energy. 324, 119775.

[17] Chouaibi, M., Daoued, K.B., Riguane, K., Rouissi, T., Ferrari, G., 2020. Production of bioethanol from pumpkin peel wastes: comparison between response surface methodology (RSM) and artificial neural networks (ANN). Industrial Crops and Products. 155, 112822.

[18] Coşgun, A., Günay, M.E., Yıldırım, R., 2021. Exploring the critical factors of algal biomass and lipid production for renewable fuel production by machine learning. Renewable Energy. 163, 1299-1317.

[19] Coşgun, A., Günay, M.E., Yıldırım, R., 2022. Analysis of lipid production from Yarrowia lipolytica for renewable fuel production by machine learning. Fuel. 315, 122817.

[20] Culaba, A.B., Mayol, A.P., San Juan, J.L.G., Vinoya, C.L., Concepcion, R.S., 2nd, Bandala, A.A., Vicerra, R.R.P., Ubando, A.T., Chen, W.H., Chang, J.S., 2022. Smart sustainable biorefineries for lignocellulosic biomass. Bioresour. Technol. 344(Part B), 126215.

[21] Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.H., Nelson, L.J., Hart, G.L.W., Sanvito, S., Buongiorno-Nardelli, M., Mingo, N., Levy, O., 2012. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput *ab initio* calculations. Comput. Mater. Sci. 58, 227-235.

[22] Das, S., Bhattacharya, A., Ganguly, A., Dey, A., Chatterjee, P.K., 2016. Artificial neural network modelling of xylose yield from water hyacinth by dilute sulphuric acid hydrolysis for ethanol production. Int. J. Environ. Technol. Manage. 19(2), 150-166.

[23] Dave, N., Varadavenkatesan, T., Selvaraj, R., Vinayagam, R., 2021. Modelling of fermentative bioethanol production from indigenous *Ulva prolifera* biomass by *Saccharomyces cerevisiae* NFCCI1248 using an integrated ANN-GA approach. Sci. Total Environ. 791, 148429.

[24] Dharmalingam, B., Tantayotai, P., Panakkal, E.J., Cheenkachorn, K., Kirdponpattara, S., Gundupalli, M.P., Cheng, Y.S., Sriariyanun, M., 2022. Organic Acid Pretreatments and Optimization Techniques for Mixed Vegetable Waste Biomass Conversion into Biofuel Production. BioEnergy Res.

[25] Dong, C., Chen, J., 2019. Optimization of process parameters for anaerobic fermentation of corn stalk based on least squares support vector machine. Bioresour. Technol. 271, 174-181.

[26] Erdem Günay, M., Yıldırım, R., 2020. Recent advances in knowledge discovery for heterogeneous catalysis using machine learning. Catal. Rev. 63(1), 120-164.

[27] Ethaib, S., Omar, R., Mazlina, M.K.S., Radiah, A.B.D., Syafiie, S., 2016. Development of a hybrid PSO-ANN model for estimating glucose and xylose yields for microwave-assisted pretreatment and the enzymatic hydrolysis of lignocellulosic biomass. Neural Comput. Appl. 30(4), 1111-1121.

[28] Ezzatzadegan, L., Yusof, R., Morad, N.A., Shabanzadeh, P., Muda, N.S., Borhani, T.N., 2021. Experimental and artificial intelligence modelling study of oil palm trunk sap fermentation. Energies. 14(8), 2137.

[29] Feng, S., Zhou, H., Dong, H., 2019. Using deep neural network with small dataset to predict material defects. Mater. Des. 162, 300-310.

[30] Fischer, J., Lopes, V.S., Cardoso, S.L., Coutinho Filho, U., Cardoso, V.L., 2017. Machine learning techniques applied to lignocellulosic ethanol in simultaneous hydrolysis and fermentation. Braz. J. Chem. Eng. 34(1), 53-63.

[31] Franco, B.M., Navas, L.M., Gómez, C., Sepúlveda, C., Acién, F.G., 2019. Monoalgal and mixed algal cultures discrimination by using an artificial neural network. Algal Res. 38, 101419.

[32] Galbe, M., Wallberg, O., 2019. Pretreatment for biorefineries: a review of common methods for efficient utilisation of lignocellulosic materials. Biotechnol. Biofuels. 12, 294.

[33] Gama, R., Van Dyk, J.S., Burton, M.H., Pletschke, B.I., 2017. Using an artificial neural network to predict the optimal conditions for enzymatic hydrolysis of apple pomace. 3 Biotech. 7, 138.

[34] Ganguly, P., Das, P., 2022. Integral approach for second-generation bio-ethanol production and wastewater treatment using peanut shell waste: yield, removal, and ANN studies. Biomass Convers. Biorefin.

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

[35] Ge, S., Shi, Y., Xia, C., Huang, Z., Manzo, M., Cai, L., Ma, H., Zhang, S., Jiang, J., Sonne, C., Lam, S.S., 2021. Progress in pyrolysis conversion of waste into value-added liquid pyro-oil, with focus on heating source and machine learning analysis. Energy Convers. Manage. 245, 114638.
[36] Giordano, P.C., Beccaria, A.J., Goicoechea, H.C., Olivieri, A.C., 2013. Optimization of the hydrolysis of lignocellulosic residues by using radial basis functions modeling and particle swarm optimization. Biochem. Eng. J. 80, 1-9.
[37] Gitifar, V., Eslamloueyan, R., Sarshar, M., 2013. Experimental study and neural network modeling of sugarcane bagasse pretreatment with H_2SO_4 and O_3 for cellulosic material conversion to sugar. Bioresour. Technol. 148, 47-52.
[38] Grahovac, J., Jokić, A., Dodić, J., Vučurović, D., Dodić, S., 2016. Modelling and prediction of bioethanol production from intermediates and byproduct of sugar beet processing using neural networks. Renewable Energy. 85, 953-958.
[39] Gražulis, S., Chateigner, D., Downs, R.T., Yokochi, A., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., Le Bail, A., 2009. Crystallography Open Database–an open-access collection of crystal structures. J. appl. Crystallogr. 42(4), 726-729.
[40] Griffin, D.W., Schultz, M.A., 2012. Fuel and chemical products from biomass syngas: a comparison of gas fermentation to thermochemical conversion routes. Environ. Prog. Sustainable Energy. 31(2), 219-224.
[41] Guo, H.N., Wu, S.B., Tian, Y.J., Zhang, J., Liu, H.T., 2021. Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: a review. Bioresour. Technol. 319, 124114.
[42] Haldar, D., Shabbirahmed, A.M., Mahanty, B., 2023. Multivariate regression and artificial neural network modelling of sugar yields from acid pretreatment and enzymatic hydrolysis of lignocellulosic biomass. Bioresour. Technol. 370, 128519.
[43] Jahanbakhshi, A., Salehi, R., 2019. Processing watermelon waste using *Saccharomyces cerevisiae* yeast and the fermentation method for bioethanol production. J. Food Process Eng. 42(7), e13283.
[44] Jain, A., Hautier, G., Ong, S.P., Persson, K., 2016. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J. Mater. Res. 31(8), 977-994.
[45] Kaya, M., Hajimirza, S., 2019. Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies. Sci. Rep. 9(1), 5034.
[46] Kim, J.H., Block, D.E., Mills, D.A., 2010. Simultaneous consumption of pentose and hexose sugars: an optimal microbial phenotype for efficient fermentation of lignocellulosic biomass. Appl. Microbiol. Biotechnol. 88, 1077-1085.
[47] Kim, J.S., Lee, Y.Y., Kim, T.H., 2016. A review on alkaline pretreatment technology for bioconversion of lignocellulosic biomass. Bioresour. Technol. 199, 42-48.
[48] Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., Wolverton, C., 2015. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comput. Mater. 1(1), 15010.
[49] Konishi, M., 2020. Bioethanol production estimated from volatile compositions in hydrolysates of lignocellulosic biomass by deep learning. J. Biosci. Bioeng. 129(6), 723-729.
[50] Kucharska, K., Hołowacz, I., Konopacka-Łyskawa, D., Rybarczyk, P., Kamiński, M., 2018. Key issues in modeling and optimization of lignocellulosic biomass fermentative conversion to gaseous biofuels. Renewable Energy. 129, 384-408.
[51] Kumar, R., Tabatabaei, M., Karimi, K., Sárvári Horváth, I., 2016. Recent updates on lignocellulosic biomass derived ethanol-a review. Biofuel Res. J. 3(1), 347-356.
[52] Kumar, S., Gupta, R., Lee, Y.Y., Gupta, R.B., 2010. Cellulose pretreatment in subcritical water: effect of temperature on molecular structure and enzymatic reactivity. Bioresour. Technol. 101(4), 1337-1347.
[53] Kumar, S., Jain, S., Kumar, H., 2018. Prediction of jatropha-algae biodiesel blend oil yield with the application of artificial neural networks technique. Energy Sources, Part A. 41(11), 1285-1295.
[54] Landis, D.D., Hummelshoj, J.S., Nestorov, S., Greeley, J., Dulak, M., Bligaard, T., Norskov, J.K., Jacobsen, K.W., 2012. The Computational Materials Repository. Comput. Sci. Eng. 14(6), 51-57.
[55] Larose, D.T., Larose, C.D., 2014. Discovering Knowledge in Data: An Introduction to Data Mining, 2 (ed). Wiley, Hoboken, New Jersey.
[56] Lee, K.M., Zanil, M.F., Chan, K.K., Chin, Z.P., Liu, Y.C., Lim, S., 2020. Synergistic ultrasound-assisted organosolv pretreatment of oil palm empty fruit bunches for enhanced enzymatic saccharification: an optimization study using artificial neural networks. Biomass Bioenergy. 139, 105621.
[57] Li, H., Chen, J., Zhang, W., Zhan, H., He, C., Yang, Z., Peng, H., Leng, L., 2023. Machine-learning-aided thermochemical treatment of biomass: a review. Biofuel Res. J. 10(1), 1786-1809.
[58] Li, J., Suvarna, M., Li, L., Pan, L., Pérez-Ramírez, J., Ok, Y.S., Wang, X., 2022. A review of computational modeling techniques for wet waste valorization: Research trends and future perspectives. J. Clean. Prod. 367, 133025.
[59] Li, W., Ghosh, A., Bbosa, D., Brown, R., Wright, M.M., 2018. Comparative techno-economic, uncertainty and life cycle analysis of lignocellulosic biomass solvent liquefaction and sugar fermentation to ethanol. ACS Sustain. Chem. Eng. 6(12), 16515-16524.
[60] Liu, C.G., Liu, L.Y., Zi, L.H., Zhao, X.Q., Xu, Y.H., Bai, F.W., 2014. Assessment and regression analysis on instant catapult steam explosion pretreatment of corn stover. Bioresour. Technol. 166, 368-372.
[61] Liu, C.G., Xiao, Y., Xia, X.X., Zhao, X.Q., Peng, L., Srinophakun, P., Bai, F.W., 2019. Cellulosic ethanol production: progress, challenges and strategies for solutions. Biotechnol. Adv. 37(3), 491-504.
[62] Loow, Y.L., Wu, T.Y., Md. Jahim, J., Mohammad, A.W., Teoh, W.H., 2016. Typical conversion of lignocellulosic biomass into reducing sugars using dilute acid hydrolysis and alkaline pretreatment. Cellulose. 23, 1491-1520.
[63] Lugani, Y., Rai, R., Prabhu, A.A., Maan, P., Hans, M., Kumar, V., Kumar, S., Chandel, A.K., Sengar, R.S., 2020. Recent advances in bioethanol production from lignocelluloses: a comprehensive review with a focus on enzyme engineering and designer biocatalysts. Biofuel Res. J. 7(4), 1267-1295.
[64] Luo, H., Gao, L., Liu, Z., Shi, Y., Xie, F., Bilal, M., Yang, R., Taherzadeh, M.J., 2021. Prediction of phenolic compounds and glucose content from dilute inorganic acid pretreatment of lignocellulosic biomass using artificial neural network modeling. Bioresour. Bioprocess. 8, 134.
[65] Ma, R., Colon, Y.J., Luo, T., 2020. Transfer Learning Study of Gas Adsorption in Metal-Organic Frameworks. ACS Appl. Mater. Interfaces. 12(30), 34041-34048.
[66] Moayedi, H., Aghel, B., Foong, L.K., Bui, D.T., 2020. Feature validity during machine learning paradigms for predicting biodiesel purity. Fuel. 262, 116498.
[67] Mohaptra, S., Dash, P.K., Behera, S.S., Thatoi, H., 2016. Optimization of delignification of two *Pennisetum grass* species by NaOH pretreatment using Taguchi and ANN statistical approach. Environ. Technol. 37(8), 940-951.
[68] Mondal, P., Sadhukhan, A.K., Ganguly, A., Gupta, P., 2021. Optimization of process parameters for bio-enzymatic and enzymatic saccharification of waste broken rice for ethanol production using response surface methodology and artificial neural network-genetic algorithm. 3 Biotech. 11, 28.
[69] Moodley, P., Rorke, D.C., Kana, E.B.G., 2019. Development of artificial neural network tools for predicting sugar yields from inorganic salt-based pretreatment of lignocellulosic biomass. Bioresour. Technol. 273, 682-686.
[70] Onay, M., 2022. Sequential modelling for carbohydrate and bioethanol production from *Chlorella saccharophila* CCALA 258: a complementary experimental and theoretical approach for microalgal bioethanol production. Environ. Sci. Pollut. Res. Int. 29(10), 14316-14332.
[71] Parkhey, P., Ram, A.K., Diwan, B., Eswari, J.S., Gupta, P., 2020. Artificial neural network and response surface methodology: a

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

comparative analysis for optimizing rice straw pretreatment and saccharification. Prep. Biochem. Biotechnol. 50(8), 768-780.

[72] Pereira, R.D., Badino, A.C., Cruz, A.J.G., 2020. Framework based on artificial intelligence to increase industrial bioethanol production. Energy Fuels. 34(4), 4670-4677.

[73] Phromphithak, S., Onsree, T., Tippayawong, N., 2021. Machine learning prediction of cellulose-rich materials from biomass pretreatment with ionic liquid solvents. Bioresour. Technol. 323, 124642.

[74] Plazas Tovar, L., Ccopa Rivera, E., Pinto Mariano, A., Wolf Maciel, M.R., Maciel Filho, R., 2018. Prediction of overall glucose yield in hydrolysis of pretreated sugarcane bagasse using a single artificial neural network: good insight for process development. J. Chem. Technol. Biotechnol. 93(4), 1031-1043.

[75] Ponnuchamy, V., 2022. Multiscale modeling studies for exploring lignocellulosic biomass structure, Advanced Catalysis for Drop-in Chemicals. pp. 257-289.

[76] Pradhan, D., Jaiswal, S., Jaiswal, A.K., 2022. Artificial neural networks in valorization process modeling of lignocellulosic biomass. Biofuels, Bioprod. Biorefin. 16(6), 1849-1868.

[77] Qiao, J., Cui, H., Wang, M., Fu, X., Wang, X., Li, X., Huang, H., 2022. Integrated biorefinery approaches for the industrialization of cellulosic ethanol fuel. Bioresour. Technol. 360, 127516.

[78] Raj, T., Chandrasekhar, K., Naresh Kumar, A., Rajesh Banu, J., Yoon, J.J., Kant Bhatia, S., Yang, Y.H., Varjani, S., Kim, S.H., 2022. Recent advances in commercial biorefineries for lignocellulosic ethanol production: current status, challenges and future perspectives. Bioresour. Technol. 344(Pt B), 126292.

[79] Rashid, T., Ali Ammar Taqvi, S., Sher, F., Rubab, S., Thanabalan, M., Bilal, M., ul Islam, B., 2021. Enhanced lignin extraction and optimisation from oil palm biomass using neural network modelling. Fuel. 293, 120485.

[80] Ravindran, R., Jaiswal, A.K., 2016. A comprehensive review on pre-treatment strategy for lignocellulosic food industry waste: challenges and opportunities. Bioresour Technol. 199, 92-102.

[81] Rivera, E.C., Rabelo, S.C., dos Reis Garcia, D., Filho, R.M., da Costa, A.C., 2010. Enzymatic hydrolysis of sugarcane bagasse for bioethanol production: determining optimal enzyme loading using neural networks. J. Chem. Technol. Biotechnol. 85(7), 983-992.

[82] Scheller, H.V., Ulvskov, P., 2010. Hemicelluloses. Annu. Rev. Plant Biol. 61, 263-289.

[83] Sebayang, A.H., Masjuki, H.H., Ong, H.C., Dharma, S., Silitonga, A.S., Kusumo, F., Milano, J., 2017. Optimization of bioethanol production from sorghum grains using artificial neural networks integrated with ant colony. Ind. Crops Prod. 97, 146-155.

[84] Selvakumar, P., Kavitha, S., Sivashanmugam, P., 2018. Optimization of process parameters for efficient bioconversion of thermo-chemo pretreated *Manihot esculenta* Crantz YTP1 stem to ethanol. Waste Biomass Valorization. 10(8), 2177-2191.

[85] Sewsynker-Sukai, Y., Faloye, F., Kana, E.B.G., 2016. Artificial neural networks: an efficient tool for modelling and optimization of biofuel production (a mini review). Biotechnol. Biotechnol. Equip. 31(2), 221-235.

[86] Shet, V.B., C. Shetty, V., Siddik, A.K.G,R.,J. Shetty, N., Concepta Goveas, L., D'Mello, G., Vaman Rao, C.P,U.,A,A., 2018a. Optimization of microwave assisted H2so4 hydrolysis of cocoa pod shells: comparison between response surface methodology and artificial neural network and production of bioethanol thereof. J. Microbiol. Biotechnol. Food Sci. 7(5), 473-477.

[87] Shet, V.B., Palan, A.M., Rao, S.U., Varun, C., Aishwarya, U., Raja, S., Goveas, L.C., Vaman Rao, C., Ujwal, P., 2018b. Comparison of response surface methodology and artificial neural network to enhance the release of reducing sugars from non-edible seed cake by autoclave assisted HCl hydrolysis. 3 Biotech. 8(2), 127.

[88] Sidana, A., Yadav, S.K., 2022. Recent developments in lignocellulosic biomass pretreatment with a focus on eco-friendly, non-conventional methods. J. Clean. Prod. 335, 130286.

[89] Sindhu, R., Binod, P., Pandey, A., 2016. Biological pretreatment of lignocellulosic biomass-an overview. Bioresour. Technol. 199, 76-82.

[90] Singhal, A., Kumar, M., Bhattacharya, M., Kumari, N., Jha, P.K., Chauhan, D.K., Thakur, I.S., 2018. Pretreatment of *Leucaena leucocephala* wood by acidified glycerol: optimization, severity index and correlation analysis. Bioresour. Technol. 265, 214-223.

[91] Smuga-Kogut, M., Kogut, T., Markiewicz, R., Słowik, A., 2021. Use of machine learning methods for predicting amount of bioethanol obtained from lignocellulosic biomass with the use of ionic liquids for pretreatment. Energies. 14(1), 243.

[92] Sousa Jr, R., Carvalho, M.L., Giordano, R.L.C., Giordano, R.C., 2011. Recent trends in the modeling of cellulose hydrolysis. Braz. J. Chem. Eng. 28(4), 545-564.

[93] Suresh, T., Sivarajasekar, N., Balasubramani, K., Ahamad, T., Alam, M., Naushad, M., 2020. Process intensification and comparison of bioethanol production from food industry waste (potatoes) by ultrasonic assisted acid hydrolysis and enzymatic hydrolysis: statistical modelling and optimization. Biomass Bioenergy. 142, 105752.

[94] Suvarna, M., Jahirul, M.I., Aaron-Yeap, W.H., Augustine, C.V., Umesh, A., Rasul, M.G., Günay, M.E., Yildirim, R., Janaun, J., 2022. Predicting biodiesel properties and its optimal fatty acid profile via explainable machine learning. Renewable Energy. 189, 245-258.

[95] Taiwo, A., Madzimbamuto, T., Ojumu, T., 2018. Optimization of corn steep liquor dosage and other fermentation parameters for ethanol production by *Saccharomyces cerevisiae* type 1 and anchor instant yeast. Energies. 11(7), 1740.

[96] Vani, S., Sukumaran, R.K., Savithri, S., 2015. Prediction of sugar yields during hydrolysis of lignocellulosic biomass using artificial neural network modeling. Bioresour. Technol. 188, 128-135.

[97] Villars, P., Cenzual, K., 2007. Pearson's crystal data®: crystal structure database for inorganic compounds. ASM International Materials Park, OH.

[98] Vinitha, N., Vasudevan, J., Gopinath, K.P., 2022. Bioethanol production optimization through machine learning algorithm approach: biomass characteristics, saccharification, and fermentation conditions for enzymatic hydrolysis. Biomass Convers. Biorefin.

[99] Vollmer, N.I., Driessen, J.L.S.P., Yamakawa, C.K., Gernaey, K.V., Mussatto, S.I., Sin, G., 2022. Model development for the optimization of operational conditions of the pretreatment of wheat straw. Chem. Eng. J. 430, 133106.

[100] Wang, Z., Peng, X., Xia, A., Shah, A.A., Huang, Y., Zhu, X., Zhu, X., Liao, Q., 2022. The role of machine learning to boost the bioenergy and biofuels conversion. Bioresour. Technol. 343, 126099.

[101] Xu, Z., Huang, F., 2014. Pretreatment methods for bioethanol production. Appl. Biochem. Biotechnol. 174(1), 43-62.

[102] Yan, L., Ma, R., Li, L., Fu, J., 2016. Hot water pretreatment of lignocellulosic biomass: an effective and environmentally friendly approach to enhance biofuel production. Chem. Eng. Technol. 39(10), 1759-1770.

[103] Yang, B., Wyman, C.E., 2008. Pretreatment: the key to unlocking low-cost cellulosic ethanol. Biofuels Bioprod. Biorefin: Innovation Sustainable Econ. 2(1), 26-40.

[104] Yaoyang, X., Boeing, W.J., 2013. Mapping biofuel field: a bibliometric evaluation of research output. Renew. Sust. Energy Rev. 28, 82-91.

[105] Zhang, Y., Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. npj Comput. Mater. 4(1), 25.

[106] Zhao, F., Yi, Y., Lü, X., 2021. Chapter 4-Essential process and key barriers for converting plant biomass into biofuels, Advances in 2nd Generation of Bioethanol Production. pp. 53-70.

[107] Zheng, Y., Pan, Z., Zhang, R., 2009. Overview of biomass pretreatment for cellulosic ethanol production. Int. J. Agric. Biological Eng. 2(3), 51-68.

[108] Zhu, X., Li, Y., Wang, X., 2019. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour. Technol. 288, 121527.

[109] Zoghlami, A., Paes, G., 2019. Lignocellulosic biomass: understanding recalcitrance and predicting hydrolysis. Front Chem. 7, 874.

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5

Ahmet Coşgun is currently a Ph.D. candidate in the Department of Chemical Engineering at Boğaziçi University. In addition to his Ph.D. studies, he is a full-time employee at a petroleum refinery. His research interest includes machine learning applications in biofuel production. His Google Scholar profile can be found at the following link:
https://scholar.google.com.tr/citations?user=WxbNtlMAAAAJ&hl=en

M. Erdem Günay is a professor in the Department of Energy Systems Engineering at Istanbul Bilgi University. He has a Ph.D. degree from Boğaziçi University, Istanbul. He has published over 35 research articles that have received approximately 1000 citations and has an h-index of 18 in the academic press. His current research interests include modeling, simulation, and optimization of renewable energy resources such as solar, wind, and biomass energy systems by machine learning. His Google Scholar profile can be found at the following link:
https://scholar.google.com.tr/citations?user=QigB6GQAAAAJ&hl=en

Ramazan Yıldırım is a professor in Department of Chemical Engineering at Boğaziçi University. He has a PhD degree from University of California, Los Angeles. He has published over 60 research articles that have received approximately 2500 citations in the academic press, and he has an h-index of 27 in the academic press. His current research areas are photocatalysis, photovoltaics and machine learning applications in renewable energy fields. His Google Scholar profile can be found at the following link:
https://scholar.google.com/citations?user=xjsiM_kAAAAJ&hl=en

Please cite this article as: Coşgun A., Günay M.E., Yıldırım R. A critical review of machine learning for lignocellulosic ethanol production *via* fermentation route. Biofuel Research Journal 38 (2023) 1859-1875. DOI: 10.18331/BRJ2023.10.2.5