MACHINE LEARNING ANALYSIS OF BIOFUELS USING PUBLISHED DATA

by

Ahmet Coşgun

B.S., Chemical Engineering, Boğaziçi University, 2015

M.S., Chemical Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Chemical Engineering

Boğaziçi University

2024

ACKNOWLEDGEMENTS

First and foremost, I would like to express my very great appreciation to my thesis supervisor, Prof. Ramazan Yıldırım for his patience, motivation, and continuous support of my research. This dissertation would not have been possible without his inspiration and effort.

I would like to express my great appreciations for my dear thesis committee members, Prof. M. Erdem Günay, Assoc. Prof. Betül Uralcan Kılavuz, Prof. Ayşe Nilgün Akın and Assoc. Prof. Damla Eroğlu Pala for spending their time and their contributions to my dissertation. I am also very grateful to Prof. M. Erdem Günay, with whom I had the privilege to collaborate on the thesis.

For a total of fifteen years in Boğaziçi University as a member of Department of Chemical Engineering and as a member of and KB405 for eight years, I would like to thank my labmates, I think of that thesis as a collaborative teamwork of the entire lab and would not be that comprehensive if I did not get assistance from you.

I am enormously grateful to my friends from Boğaziçi, my hometown Ataköy, to my feloows from Tüpraş, and to many others who have supported me throughout my studies. Thank you all! I could not have made it without you.

The last acknowledgement goes to my parents, Fatmagül Coşgun and Sadrettin Coşgun, for their continuous support and motivation.

ABSTRACT

MACHINE LEARNING ANALYSIS OF BIOFUELS USING PUBLISHED DATA

One of the key objective of this dissertation is to develop a comprehensive understanding of previous machine learning (ML) studies on various biofuel production routes, specifically microalgae biodiesel, lignocellulosic ethanol, and biochar production, through bibliometric analysis and literature review. Microalgae show high potential for biofuel production but are costly, requiring better strains, value-added chemicals and reduced energy consumption for commercial viability. Converting lignocellulosic biomass to ethanol is promising but remains expensive, particularly due to pretreatment step. Biochar, especially in its functionalized and engineered forms, is likely to become more important for environmental remediation. This work also aims to construct comprehensive datasets for lipid production from the oleaginous yeast strain *Yarrowia lipolytica* and tri-reforming process from biogas, analyzing them with ML tools to estimate unperformed experiments, determine critical variables, and develop optimal heuristics. It was found that the potential of *Yarrowia lipolytica* for lipid production can be effectively categorized using decision tree (DT) algorithms, and descriptor effects can be determined using association rule mining. Found that, C/N ratio with fermentation time being influential for biomass production, while pH and glucose concentration for lipid content. The tri-reforming models developed using random forest (RF) and artificial neural network (ANN) algorithms also showed high predictive capabilities. Transfer learning methods, using source models from dry-reforming and steam reforming datasets were implemented to enhance the prediction accuracy with smaller datasets. DT algorithms were employed for interpretable knowledge extraction, resulting in the development of hierarchical rules for methane conversion, carbon dioxide conversion, and H_2/CO ratios in the produced syngas. While high reaction temperature with high W/F results with high CH_4 conversion, moderate reaction temperature with low O_2 ratio promote high CO_2 conversion. In all cases, not depending on other conditions, increasing H_2O, and decreasing CO_2 in the feed composition increases H_2/CO ratio. The findings align with the literature, indicating that data mining and machine learning methods effectively provide predictive models and detect trends and patterns that are not readily observable.

ÖZET

YAYINLANMIŞ VERİLER İLE BİYOYAKITLARIN MAKİNE ÖĞRENİMİ ANALİZİ

Bu tezin amaçlarından biri, bibliyometrik analiz ve literatür incelemesi yoluyla, farklı biyoyakıt üretim yöntemleri üzerindeki makine öğrenmesi (ML) çalışmalarını, özellikle mikroalg biyodizel, lignoselülozik etanol ve biyokömür üretimini kapsamlı bir şekilde anlamaktır. Mikroalgler, biyoyakıt üretimi için yüksek potansiyele sahip olmalarına ragmen, ticarileşebilmesi için farklı türler, katma değerli kimyasallar ve enerji tüketiminin azaltılması gerekmektedir. Lignoselülozik biyokütlenin etanole dönüştürülmesi de umut verici olmakla birlikte özellikle ön işlemler nedeniyle pahalıdır. Biyokömür, özellikle fonksiyon kazandırılmış formları, çevresel iyileştirme çalışmaları için önemlidir. Bu çalışma ayrıca, yağlı maya türü *Yarrowia lipolytica*'nın yağ üretimi ve biyogazın tri-reforming süreciyle sentez gazına dönüştürülmesi için kapsamlı veri setleri oluşturarak, bunları ML araçlarıyla analiz etmeyi, bilinmeyen koşullarda elde edilebilecek sonuçları tahmin etmeyi, kritik değişkenleri belirlemeyi amaçlamaktadır. *Yarrowia lipolytica*'nın lipid üretim potansiyelinin karar ağacı (DT) kullanılarak etkili bir şekilde kategorize edilebileceği ve değişken etkilerinin ilişkilendirme kural madenciliği kullanılarak belirlenebileceği bulunmuştur. C/N oranı ve fermantasyon süresi biyokütle üretimi üzerinde etkili bulunurken, pH ve glikoz konsantrasyonu lipid içeriği için önemlidir. Rastgele orman (RF) ve yapay sinir ağı (ANN) kullanılarak geliştirilen tri-reformlama modelleri yüksek tahmin yeteneği göstermiştir. Küçük veri setlerinde tahmin doğruluğunu arttırmak için buhar ve CO_2 reformlama verilerini kullanan transfer öğrenme yöntemi uygulanmıştır. Yorumlanabilir bilgi çıkarımı için DT algoritmaları kullanılmış ve metan dönüşümü, karbondioksit dönüşümü ve H_2/CO oranları için genellenebilir kurallar geliştirilmiştir. Yüksek reaksiyon sıcaklığı ve yüksek W/F, yüksek CH_4 dönüşümü sağlarken, orta reaksiyon sıcaklığı ve düşük O_2 oranı, yüksek CO_2 dönüşümü sağlamıştır. Tüm durumlarda, diğer koşullardan bağımsız olarak, besleme bileşimindeki H_2O'nun artırılması ve CO_2'nin azaltılması H_2/CO oranını artırmaktadır. Bulgular, veri madenciliği ve makine öğrenme yöntemlerinin etkili öngörü modelleri sağladığını ve kolayca gözlemlenmeyen eğilimleri tespit ettiğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AB	Adaptive Boosting
AD	Anaerobic Digestion
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
ASH	Ash content (%)
BOD	Biochemical Oxygen Demand
C-char	Carbon contents of biochar
CEF	Cellulose Enrichment Factor
CEL	Cellulose (%)
CNN	Convolutional Neural Network
COD	Chemical Oxygen Demand
CSF	Combined Severity Factor
DA	Dilute-Acid Treatment
DES	Deep Eutectic Solvent
DT	Decision Tree
FC	Fixed carbon (%)
GB	Gradient Boosting
GP	Genetic Programming
HC	Hemicellulose (%)
HHV	Higher heating value (MJ/kg)
HR	Heating rate (°C/min)
HTC	Hydrothermal Carbonization
HTL	Hydrothermal Liquefaction
HTT	Highest treatment temperature
IL	Ionic liquids
k-NN	k-Nearest Neighbor
KRR	Kernel Ridge Regression
LCB	Lignocellulosic Biomass
LCW	Lignocellulosic Waste

LIG	Lignin (%)
LogR	Logistic Regression
LSTM	Long Short-Term Memory Network
MARS	Multivariate Adaptive Regression Splines
MC	Moisture Content (%)
ML	Machine learning
MLP	Multilayer Perceptron
MLR	Multi-Linear Regression
MSW	Municipal Solid Waste
NB	Naive Bayes
N-char	Nitrogen contents of biochar
OD-MS	Optimized Decision-Making System
OM	Organic matter
PCA	Principal Component Analysis
PRM	Polynomial Regression Model
PS	Particle Size of Biomass
PSO	Particle Swarm Optimization
QDA	Quadratic Discriminant Analysis
RBF	Radial Basis Functions
RF	Random Forest
RidReg	Ridge Regression
RSM	Response Surface Methodology
RT	Retention time
SCB	Sugarcane Bagasse
SE	Saccharification Efficiency
SHF	Separate Hydrolysis and Fermentation
SR	Solid Recovery
SSA	Specific surface area (m^2/g)
SSCF	Simultaneous Saccharification and co-Fermentation
SSF	Simultaneous Saccharification and Fermentation
STE	Steam Explosion
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting Machine

1. INTRODUCTION

Biofuels are essential to meet the sustainable energy demands of the future. According to the International Renewable Energy Agency (IRENA), while transportation will become more electrified, biofuels will remain essential for several decades, particularly for sectors that cannot easily transition to electric vehicles (EVs). EVs, powered increasingly by renewable electricity, will dominate light vehicle fleets in regions with well-developed power grids and charging infrastructure. However, biofuels will be indispensable for decarbonizing heavy long-distance freight trucks, marine ships, and airplanes, which require higher energy densities and are unlikely to be fully electrified. Two main types of fuels exist: those made from fossil resources and those from renewable resources, including biological materials. Biomass can be converted into various forms of energy and value-added products through thermochemical and biological processes. Biofuels encompass solid, liquid, and gaseous fuels. Advanced biofuels using lignocellulosic feedstocks, waste, and microorganisms could significantly expand the resources for energy need in the world. The choice of conversion process depends on the quantity and type of biomass and the form of energy required.

The application of machine learning (ML) techniques in biofuel production shows significant promise, given the complexity of biochemical processes. In the past decade, there has been a shift from physical modeling to data-driven modeling in the biochemical industry. The complexity of bioprocesses, which is often due to poorly defined biochemical systems, complex raw materials, lack of understanding of biological mechanisms, genetic and environmental variations, and the multiphase nature of bioreactor systems, makes the first principal modeling quite difficult. In such cases, with the ever-increasing accumulation of digital data in the wield, the machine learning techniques can provide predictive models and heuristics rules for future studies while they can contribute to the better understanding of biofuel systems.

In this dissertation, the use of machine learning in various biofuel systems were investigated. While the field of biofuels are too wide and complex to cover in a single dissertation, the level of ML use reported in literature also differs among the individual

biofuel technologies. Hence, bibliometric analysis, with an extensive literature review, were performed for some biofuel technologies (algal biodiesel, lignocellulosic bioethanol and biochar), and detailed ML analysis were performed for others (oleaginous yeasts, and biogas reforming processes) by constructing extensive data sets from the literature and using them to develop predictive models, as well as heuristic rules to understand the effects of various descriptors on the relevant target process variables.

In Section 2, a brief information about biofuels is given (Section 2.1), with a detailed explanation of routes to produce biodiesel from microalgae and oleaginous yeasts, bioethanol from lignocellulosic biomass, biochar, and tri-reforming process for syngas production from biogas in individual sections, together with a brief information about machine learning (Section 2.2). Then, a comprehensive review about machine learning algorithms employed in the scientific articles together with keyword analysis are summarized in Section 3, where examples within the scope of microalgae biodiesel production is given in Section 3.1, lignocellulosic bio-ethanol production via fermentation is given in Section 3.2, and biochar production is presented in Section 3.3. In Section 4, all detailed procedures for analyses together with the results and discussions gathered for oleaginous yeast is presented. First, the result of preliminary analysis is presented (Section 4.1), then results and discussions from oleaginous yeast dataset is presented (Section 4.2); as computational details provided in Section 4.2.1, result of decision tree (DT) algorithm in Section 4.2.2, and association rule mining (ARM) algorithm in Section 4.2.3. Section 5 gives the procedures for analyses together with the results and discussions gathered for biogas tri-reforming. In Section 5.1, variables effecting the process is discussed thoroughly. In Section 5.2 results and discussions from biogas tri-reforming process is presented with computational details in Section 5.2.1, result of artificial neural network (ANN) in Section 5.2.2, random forest (RF) in Section 5.2.3, and DT in 5.2.4. In Section 5.3, computational details of transfer learning (TL) methodology are presented in Section 5.3.1, followed by the results of TL methodology are shared for CH_4 conversion (Section 5.3.2), and CO_2 conversion and H_2/CO ratio (Section 5.3.3). Section 6 summarizes the major conclusions and recommendation for the future studies.

2. BACKGROUND

2.1. Biofuels

Biofuels refer to energy-rich chemicals derived from the biomass of living organisms, such as microalgae, plants, and waste. The increasing need for viable and environmentally friendly energy supplies to improve the quality of life makes biofuels a crucial source to meet this demand. Although fossil fuels have long been the primary energy source, their use is unsustainable and leads to environmental issues related to combustion. The main advantage of using biomass as an energy resource is its renewability and non-contribution to global warming [1].

The conversion of biomass into energy is achieved through three primary process technologies: biochemical, thermochemical, and physicochemical. Biochemical conversion includes anaerobic digestion, which converts biomass into biogas, and fermentation, which produces ethanol. Thermochemical conversion involves four main processes: torrefaction, pyrolysis, gasification, and liquefaction. Physicochemical conversion primarily involves extracting the inner lipid content of oleaginous microorganisms and oilseeds, which are then converted into biodiesel through the transesterification process [2]. Figure 2.1 shows these conversion methodologies; it is also obvious from the figure, the biomass used as the source of the process may vary from industrial, agriculture and domestic waste to lignocellulosic materials and microorganisms.

As discussed in Introduction, investigating all the processes and feedstock used in biofuel production within the context of a dissertation is impossible (and also unnecessary considering the works have been already reported in the literature), only some of some of these processes that draw more attention in recent years have been investigated; hence, the literature survey was also restricted to biodiesel derived from microalgae (Section 2.1.1), bioethanol derived from lignocellulosic feedstock via fermentation route (Section 2.1.2), biochar production with thermochemical processes (Section 2.1.3), biodiesel derived from oleaginous yeasts (Section 2.1.4), and biogas tri-reforming process to syngas production (Section 2.1.5) are discussed in detail, as they are the focus of this dissertation.

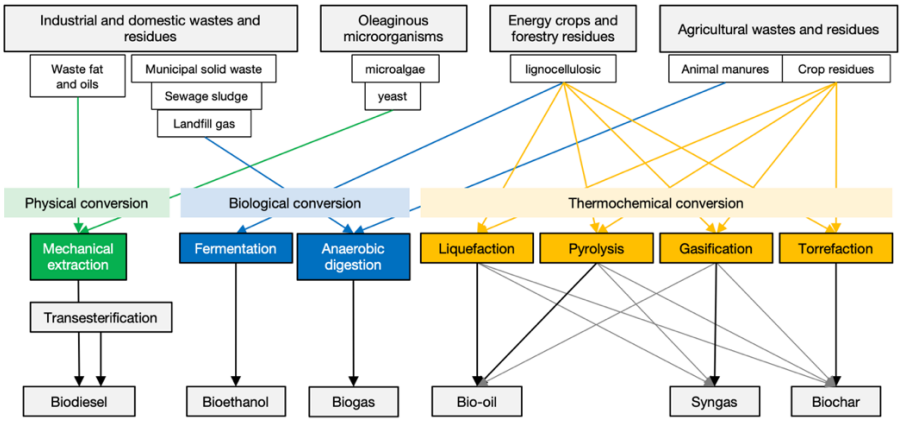

Figure 2.1. Main processes for biomass to energy conversion

2.1.1. Microalgae Biodiesel

As a new-generation feedstock for the production of biofuels, microalgae offer tremendous potential due to their rapid growth and flexibility of cultivation on terrains that are not suitable for agriculture. Microalgae are one of the microorganisms that can be utilized to produce a wide range of biofuels and bioproducts in a sustainable manner [3]. While the photosynthetic pathway for the production of sustainable bioproducts makes the microalgae ideal agents for CO_2 emission mitigation, their ability to grow in even polluted waters makes them useful for wastewater treatment [4, 5]. Moreover, microalgae can be used to produce biofuel [3, 6], nutritious food supplements and animal food [7] and can also be used for wastewater treatment and CO_2 sequestration [8]. In addition, high-value bioproducts that find widespread application in the pharmaceutical and cosmetic industries can be manufactured [9]. However, the growth of algae involves numerous stages and requires the optimization of a vast number of variables such as light-related parameters, temperature, pH and nutrients. In addition to cultivation and biomass conversion to biofuels, the steps like harvesting, drying, and extraction need to be also improved for economically feasible algal biofuel production. The feasibility can be improved further by producing value-added co-products as the popular *biorefinery strategy* implies. As a result, a number of obstacles must be overcome for the successful commercialization of this potentially important technology [10]. In addition to inevitable experimental works to be performed, computer-aided technologies can contribute to this as they are used in various bio-production industries, including strain selection and the optimization of different production steps [11, 12].

Algae can grow in a variety of habitats, including open and closed ponds, photobioreactors, sewage and wastewater, desert, marine and seawater, and CO_2-generating power plants [13]. Microalgal biomass is primarily made up of carbohydrates and lipids and is generated by photosynthesis with carbon dioxide and nutrients like nitrogen, phosphorus and potassium [14]. These lipid and carbohydrate components are raw materials that can be utilized in the production of a wide variety of biofuels as shown in Figure 2.2. Microalgae can be converted into biofuels by either biochemical or thermochemical processes depending on the type of energy or product to be produced as well as economic considerations [15]. The triglyceride (*i.e.*, lipid) component of the algae can be converted to fatty acid esters (biodiesel) and glycerin in the presence of the addition of a short-chain alcohol and a catalyst by the transesterification process [16]. After lipid extraction carbohydrate content can be turned into biogas by anaerobic digestion [17], and cellulose or starch content can be transformed into bioethanol by fermentation [18]. The pyrolysis process, in which biomass is thermally decomposed in the absence of oxygen, yields char, bio-oil, and other gas products [19]. Bio-hydrogen can be produced as it can produce hydrogen by fixing photosynthetic CO_2 in the presence of sunlight and subsequently fermenting and degrading stored organic molecules [20]. In addition to biofuel production, the algal biomass contains molecules that can be converted to a variety of value-added products as food, cosmetics, pharmaceutical additives, fertilizers and other chemicals [19].

Biofuel production from microalgae involves a number of stages, including cultivation, harvesting, drying, lipid extraction, and conversion to biofuels; the biodiesel production is presented in Figure 2.3. Growing algal strains with high lipid productivity is crucial for industrial applications. Apart from the selection of suitable strain, the cultivation of microalgae is affected by a number of factors like reactor type, growth method, properties of light, carbon source, nitrogen source, phosphorus source and operating conditions (*e.g.*, temperature, pH and salinity) as shown in Table 2.1 [21, 22]. Increasing biomass and lipid product yield by optimizing microalgae growth conditions is a practical strategy that have been adopted frequently. In the majority of instances, increasing the concentration of nutrients and CO_2 results in an increase in biomass productivity up to a certain level; however, exceeding this level causes stress, which in turn results in a reduction in growth and can even lead to the death of the cell [21]. On the other hand, microalgae tend to accumulate lipids when grown under stressful environments [23] creating a difficult

optimization problem to solve; the lipid content of fast-growing algae tends to be very low, whereas the lipid contents of slow-growing species is generally high. For this reason, the environmental stress is applied as a strategy to stimulate lipid production [24]. Chemical stress (*i.e.*, nutrient deficiency, salinity, pH) and physical stress (*i.e.*, temperature, light intensity) are two common types of environmental stress applied [21]. To improve lipid production, a two-stage cultivation procedure is frequently used; first, maximum biomass productivity is reached, and then a stress environment is generated to favor lipid synthesis.

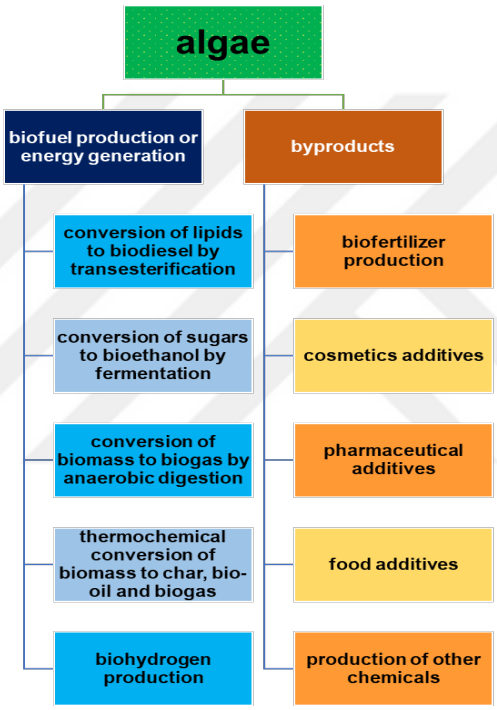

Figure 2.2. Biofuels and coproducts than can be produced from algal biomass [25].

Before proceeding with the extraction of lipids, the microalgal biomass must be harvested and dried to increase biomass concentration. However, microalgae are in the range of 3-30 micrometers in diameter making the cost-effective and efficient harvesting difficult [26, 27]. Some methods used in harvesting are screening, flocculation, filtration, gravity sedimentation, flotation and centrifugation [28]. Although centrifugation and microfiltration are utilized in the majority of laboratories, their high energy and capital costs make them infeasible for large-scale procedures [27, 29, 30]. Gravity sedimentation and flocculation are the most common ways to harvest on a large scale because they are easy to implement and have lower costs; however, they are not very effective, and they may cause biomass deterioration [31]. Then, the biomass is dried to reduce intracellular water content, which

creates mass transfer barrier during the following lipid extraction step. Although the lipid content does not vary with the method of drying [32, 33], the high temperatures attained in drying operations adversely affect the lipid quality [34]. In freeze-drying, pressure is used to reduce the water content without altering the quality of the lipids, and it is one of the most frequently used techniques on a laboratory scale. The process of drying demands a substantial amount of energy, which accounts for the majority of the total amount of energy required for the synthesis of biofuel from microalgae [35, 36]. Although solar drying is one of the most cost-effective approaches, it is also significantly slower compared to oven and freeze-drying processes. There are also published studies on the direct extraction of lipids from wet microalgae for more cost-effective biofuel manufacturing [37].

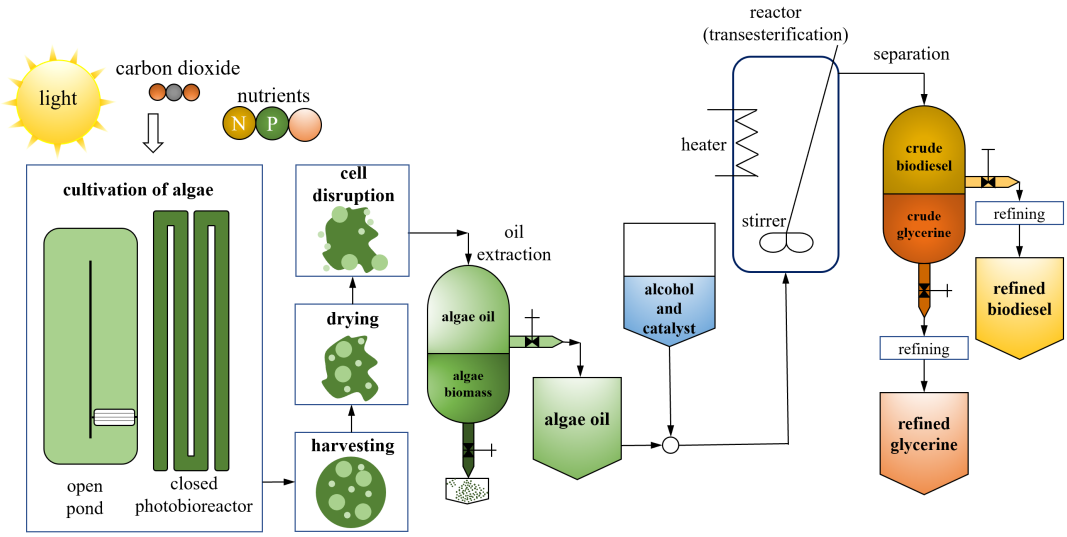

Figure 2.3. Process diagram of biodiesel production from microalgae [21].

The common methods used for lipid extraction can be categorized as solvent accelerated fluid and supercritical fluid extraction [38]. The primary consideration in extraction is to avoid environmentally harmful solvents such as chloroform and methanol [39] and use less harmful alternatives like ethyl acetate and CO_2 (in supercritical fluid extraction) [40-43]. Also, since the lipid contents of most microalgae species are not too high (*i.e.*, 15-30%), lipid must be recovered almost completely to be economical [42]. Furthermore, the solvent should be selective to desired lipid fractions [44-46]. Pretreatment methods such as mechanical or non-mechanical cell disruption prior to lipid extraction can also be used to improve the efficiency of the extraction process [47].

Table 2.1. Main conditions to be optimized for algae growth.

Variables	**Properties**
reactor type	raceway open pond, flat-plate type photobioreactor, inclined tubular type photobioreactor, horizontal/continuous type photobioreactor
growth method	photoautotrophic, heterotrophic, mixotrophic
light related variables	light intensity, light wavelength, photoperiod time
carbon source	inorganic (CO_2 or carbonates) or organic (sugars)
nitrogen source	nitrate (NO_3^-), nitrite (NO_2^-), ammonium (NH_4^+), urea
phosphorus source	polyphosphates, orthophosphates, organic phosphates
operating conditions	temperature, pH, salinity, flow rate

Biofuel production from microalgae appears to be one of the best renewable energy alternatives because microalgae can grow at a faster rate than most of the other bioresources and it does not compete with food chain. However, the potential of algae biofuels, as replacement for fossil fuels, is not that obvious yet; although the works reporting techno-economic assessment of algal biofuels are usually optimistic [48, 49], they rely on various (and often dissimilar) assumptions because the technology and market are not established yet. For instance, the report by Klein and Davis from NREL, based on the data from Arizona Center for Algae Technology and Innovation (open pond), estimated minimum fuel selling price of $5.90 gasoline equivalent for the fuel obtained [50]. On the other hand, Xin *et al.* reported that the selling price of bio-oil would be $1.85/gallon with the payback periods of 9.05 to 7.47 years assuming an integrated system having biofuel production, wastewater treatment improvement, tax credits, carbon credit, and co-products utilization [51]. In another recent work, Hassain *et al.* calculated the price of $2.51/gallon with the payback period of 11 years [52].

Despite the existing uncertainties, the technology is progressing fast as it was also indicated by the NREL report mentioned above; significant improvements in almost all cost items as well as yield and annual productivity are projected for the year 2030. For example, ash-free dry weight (AFDW) increased from 12.4 ton AFDW/acre-year in 2015 to 25.5 AFDW/acre-year in 2021 and expected to grow to 37.2 AFDW/acre-year in 2030; similarly, the annual average cultivation productivity ($g/m^2/day$) was 8.5 and 17.6 for 2015 and 2021 respectively while it was projected as 25 ($g/m^2/day$) for 2030 [50]. The progress may be

speed up in future with the new genetic modifications as Subramanian and Sayre argued that it is possible to have three fold improvement in algal biomass yield by genetic modification of wild strains [53].

Significant challenges that exist in almost all steps like strain isolation, nutrient sources, harvesting, fuel extraction and conversion has to be resolved [54]. Producing low-cost algal biofuel is difficult with the current state of art because great deal of energy and resources, including water, are needed for cultivating algae, recirculating the nutrients and drying the biomass [55]. For example, microalgae can grow in both *open ponds* and *close reactors* with their own benefits and drawbacks. Closed systems consume more energy and are costlier to make; on the other hand, open systems are relatively affordable but they are vulnerable to uncontrolled contamination and fluctuations in light, nutrients, salt and temperature [56]. Similarly, there are various options for harvesting and conversion of algal biomass to be selected and optimized. In addition to the improvements that can be made in individual steps, two major courses of action, which are highly related, are frequently suggested through the literature for the successful commercialization of algal biofuels: *selection and optimization of microalgal strain* and *production of other chemicals* in addition to the biofuel. It was argued that there are more than 30,000 known microalgae species, and less than a hundred have been studied [57]; the total number of microalgae strains changes depending on the sources and reaches to the estimated value over a million [58]. Each microalgae have different characteristics and can synthesize different products in different yields depending on process conditions. Use of one strain instead of others or utilization of multiple strains, may be more advantageous making the selection and identification of the culture composition a very important task [5]. Not only the lipid and carbohydrate content of different strains but also their growth may be different in different cultivation conditions. For instance, Nagarajan *et al.* argued that differences in COD/BOD, nitrogen and phosphorus levels in different wastewater as well as the toxic effects of ammonia, anti-inflammatory agents, antibiotics, anti-depressants and anti-epileptics in pharmaceutical waste and heavy metals such as lead, copper, chromium, mercury and zinc have to be considered, and the microalgae strain suitable for the wastewater must be chosen accordingly [59]. Production of additional chemicals, as also related to popular *biorefinery* concept, also require the judicious selection of algal strain complicating the problem more.

2.1.2. Lignocellulosic Bioethanol

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, as shown in Figure 2.4 [60]. Although the conversion of simple sugar or starch-based biomass to ethanol through fermentation is quite easy and 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 [61, 62]. Moreover, it is the most plentiful raw material on Earth that can be used to produce biofuels.

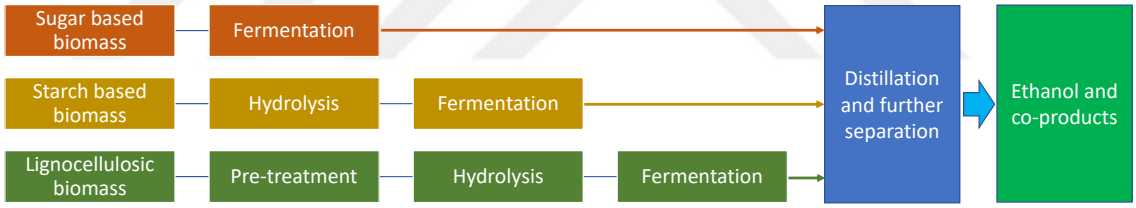

Figure 2.4. Fermentation-based transformation of different feedstocks into ethanol [63].

There are different technological pathways to produce ethanol from lignocellulosic biomass as shown in Figure 2.5, and as discussed in detail elsewhere [64]. 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; however, the syngas can also be fermented to ethanol via anaerobic digestion [65]. 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 [66]. However, the predominant technological pathway for bioethanol production is direct fermentation.

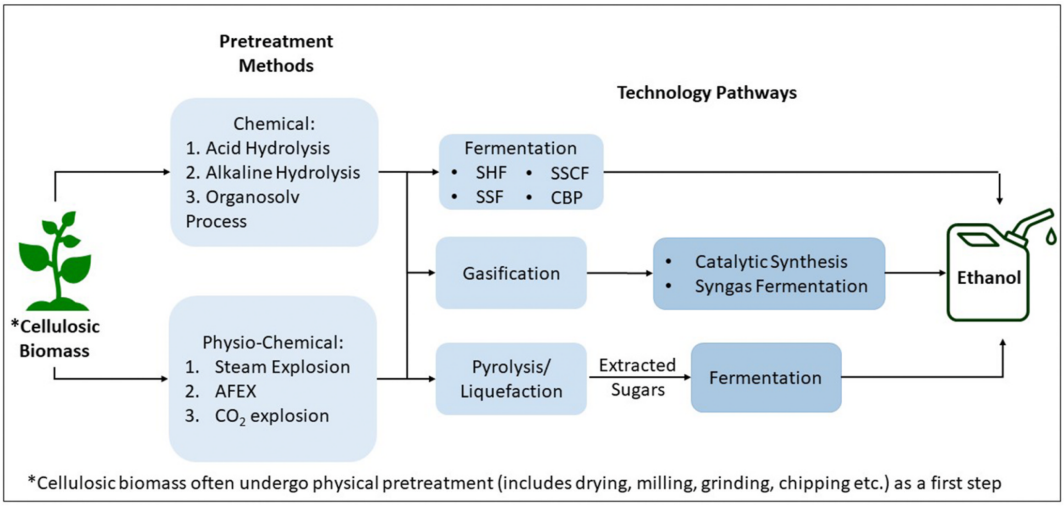

Figure 2.5. Cellulosic ethanol conversion pathways and typical pretreatment methods, reproduced from [64].

Compared to sugar and starch-based feedstock, lignocellulosic biomass is more complex, and understanding its structure, especially at a molecular level is critical [62]. It mainly consists of cellulose, hemicellulose, and lignin; all of which are tangled up in one another to make lignin-carbohydrate complexes. Cellulose is the primary component of plant cell walls giving them rigidity and strength, and it is the largest carbohydrate in the lignocellulosic biomass accounting for 40-60% of the weight. Hemicelluloses are the second most abundant carbohydrate in lignocellulosic biomass (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) [60]. Finally, lignin, as the remaining part, is an aromatic, water-insoluble polymer that provides the plant with water-proofing ability, structural strength, and resilience [67]. Compared to cellulose and hemicellulose, lignin (the protective structure) is especially resistant to biological breakdown.

Unfortunately, due to the complex nature of the lignocellulosic biomass, 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 lignocellulosic biomass are bonded together 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 [68]. Hence, pretreatment is required to separate the lignin

and recover cellulose and hemicellulose for conversion to ethanol [69]. The separated lignin can be combusted to generate heat or sold in the market [64]. 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 [70]. The pretreatment of lignocellulosic biomass is the most expensive process in the pathway, accounting for approximately 20% of the total cost [71]; 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.

There are several types of pretreatment processes 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 lignocellulosic biomass; the ultra-fine milling process can be also used to reduce cellulose polymerization and crystallinity although it is costly [62]. 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 [72]. Additionally, liquid hot water pretreatment [73] and steam explosion [74] 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 [75]. 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 [76]. However, this process also has several disadvantages such as the need for expensive corrosion-resistant equipment or the neutralization of acidic prehydrolyzates before the fermentation of sugars [77].

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 sodium hydroxide (NaOH) or ammonium hydroxide (NH_4OH) solutions. Some of the alkaline pretreatment techniques can also allow the recovery and reuse of chemical reagents [78]. 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 [77]. 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 it is not effectively removed [62]. Ionic liquids (high amount of organic cation with a small amount of inorganic anion), deep-eutectic solvents (mixtures of Lewis and Bronsted acids and bases), organosolv (organic solvents; *e.g.*, ethanol, methanol, butanol, acetone) methods can also be used [79].

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 be used to degrade lignin and hemicelluloses from lignocellulosic biomass [80]. 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 [62]. 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) [81, 82].

Following the pretreatment phase, the hydrolysis process takes place, which can be classified as acid and enzymatic hydrolysis [83]. Enzymatic hydrolysis has a lower environmental impact and inhibitor formation, while acid hydrolysis is faster [84]. In enzymatic hydrolysis, due to the fact that lignocellulosic biomass 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 are required [85]. 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 [86]. 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, the acetic acid is also produced as a byproduct of the hydrolysis of hemicellulose, and limits the microbial development and ethanol production; this can be considered as the major disadvantage of hydrolysis process [87]. Another disadvantage is the high energy consumption 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 [62].

At the end of the hydrolysis process, a soup of hexose and pentose sugars is produced. 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 [88]. Microbes that naturally ferment all of these sugars also have a low tolerance for bioreactor 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 [89]. For example, *Saccharomyces cerevisiae* is one of the most commonly used microorganisms in the fermentation process that cannot naturally utilize xylose [90]. Even though those microorganisms may utilize pentose sugars, the glucose generated from cellulose often inhibits the catabolism of these sugars [86]. All these make it difficult to develop and control the fermentation of lignocellulosic biomass 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 lignocellulosic biomass characteristics, are important for the production of bioethanol at competitive prices [91]. High efficiency, low cost and low level of inhibitory byproducts using greener pretreatment solutions are among the current research focuses [92]. Fermentative microorganisms are also in need in the field requiring more research for discovering efficient and robust microbial consortiums or the construction of genetically engineered strains [93]. 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 [94]. By taking all factors into consideration, it becomes hard

to find the optimal solution for bioethanol production for different regions of the world. The use of traditional mathematical models is a solution for this optimization problem with first-order equations [95]. 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, if more research and experimental effort is given to produce high-quality reproducible data [96].

2.1.3. Biochar Production

Lignocellulosic biomass is a critical resource to meet the continuously growing fuel demand of the world in a sustainable manner. Unlike first-generation crops like corn, sugarcane and other allied crops for the production of biofuels, lignocellulosic biomass (as well as energy crops like miscanthus and jatropha) does not compete with the food chain. Moreover, it is available globally in the form of agricultural wastes or forest residues and these residues are accessible in large quantities even from areas unsuitable for agriculture [97]. However, fuel production from lignocellulosic biomass is much more complicated than the traditional first-generation technologies like the fermentation of sugar to form bioethanol or transesterification of oils/fats to form biodiesel [98].

In prehistoric times, wood-based biomass served as the primary fuel source, and processing it to make other fuels significantly impacted human history. The Stone Age persisted for an extended period because the temperatures required to melt most metals and create metal tools were unattainable with regular wood fires, which burn at around 600°C, depending on the wood type. This changed with the discovery of how to make charcoal from biomass, enabling higher temperatures to be reached. This allowed for the melting of copper and tin to produce bronze, marking the end of the Stone Age and the beginning of the Bronze Age [99]. Charcoal remains as a major energy source for millions globally produced from various raw materials and is expected to continue in this role for the foreseeable future [100, 101]. Modern biochar production techniques can generally be applied through two main pathways: biochemical and thermochemical conversion [102]. Biochemical conversion is more cost-effective and less harmful to the environment, but it can take hours or even days to complete even in the presence of biocatalysts [103]. Additional disadvantages of this route

include complicated pretreatment processes, high enzyme costs, and relatively low conversions, while produced biochar may not have the desired properties [100]. Thermochemical processing, on the other hand, allows for the rapid processing of a wide range of feedstock and the production of fuels, chemicals, and power [104].

The thermochemical pathway for biochar production includes techniques such as pyrolysis, torrefaction, gasification, and hydrothermal carbonization as shown in Figure 2.6 [105]. For instance, pyrolysis is the decomposition of biomass in the absence of oxygen at a heating rate depending on the pyrolysis type [104]. Torrefaction, also known as low-temperature pyrolysis, produces a solid product named *torrefied biomass* in an inert or nitrogen atmosphere at low temperatures as 200-300 ºC [106]. Hydrothermal carbonization, also known as wet pyrolysis, is the carbonization process at low temperatures and pressures (typically 200-300 ºC and 2-5 MPa) with the solid product named *hydrochar* [107]. Finally, gasification is an advanced procedure that can be employed in a variety of atmospheric conditions (such as oxygen/air, steam, H_2, or CO_2) to convert biomass into biogas [108]. Pyrolysis, which is the common process for biochar production is discussed, as an example to show the input and output relations for ML; those are similar in other methods as well. In general, the type of biomass, heating rate and methodology, temperature, pressure, residence time, and pyrolysis method influence the biochar properties. In addition, catalytic pyrolysis and co-pyrolysis are also interesting techniques worth mentioning.

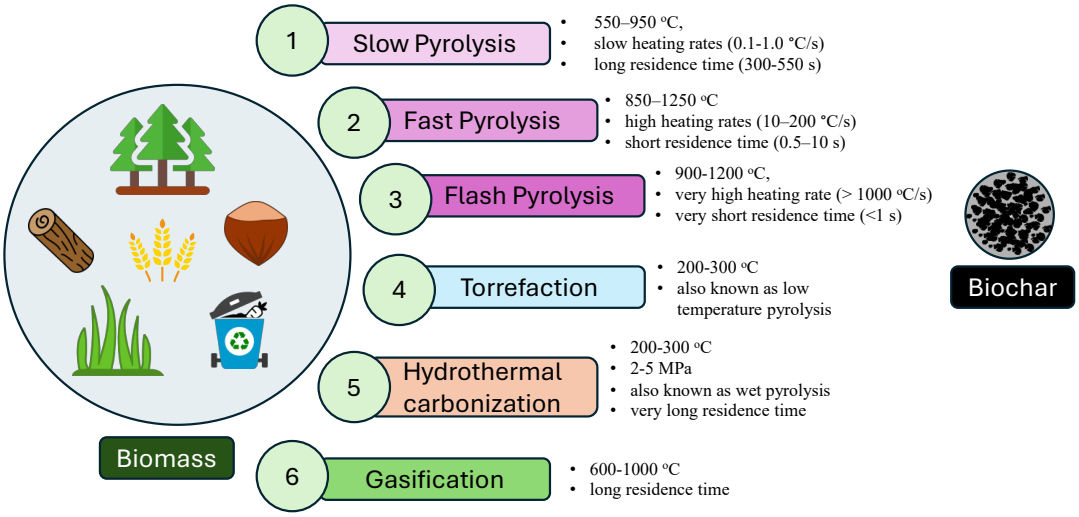

Figure 2.6. Thermochemical conversion strategies and their important characteristics.

Pyrolysis process is capable of producing fuels with different phases, as biochar, bio-oil, and gaseous products and enables use of various feedstocks [109]. Pyrolysis includes numerous reactions across gas, liquid, and solid phases and produce hundreds of volatile compounds, which make it highly intricate [110]. A typical pyrolysis process includes a reactor, cyclone and a condenser unit. The reactor heats the biomass, causing pyrolytic vapor to ascend while leaving biochar behind. At this stage, the vapor produced through pyrolysis is rather impure and contains solid particulates. Next, the gas is sent to the cyclone, where the impurities are pushed to the walls of the cyclone by the centripetal force caused by the rapid rotation of the cyclone. In the final stage, the cleaned gas passes through the condenser to produce biogas and liquid bio-oil.

Pyrolysis can be categorized into three different classes as slow, fast, and flash pyrolysis, each with different advantages and disadvantages [111]. For instance, slow pyrolysis is a more conventional pyrolysis type performed at 550–950 ºC, where slow heating rates (0.1-1.0 °C/s) and long residence times (300-550 s) are main characteristics, yielding mostly biochar and a minimal amount of liquid product. On the other hand, fast pyrolysis is performed at 850–1250 ºC with high heating rates (10–200 °C/s) and short residence times (0.5–10 s)yielding mostly liquid products and a minimal amount of biochar and biogas [112]. Flash pyrolysis, on the other hand, performed at similar reaction temperatures as fast pyrolysis, however at significantly higher heating rate (> 1000 ºC/s) and shorter residence time (<1 s) yields high liquid and lower char yield [103]. The valuable products, namely solid biochar and liquid biofuels are conversely related; enhancement of one yield results in a decrease in the other, and can be controlled with the manipulation of the process conditions [113].

Biochar is mainly produced from energy crops and forest/crop residues. However, a significant amount of work also focuses on alternative feedstocks for biochar production such as wastewater sludge, municipal waste, animal manure, and algae to achieve two tasks in one action, *i.e.*, waste mitigation and production of valuable biofuels and biochemicals [114, 115]. These feedstocks can also be advantageous concerning the utilization area of the produced biochar. Animal manure-based biochar is more suitable for carbon sequestration in the soil as it has a higher nutrition level for plants and higher cation exchange capacity compared to plant-based biochar [116]. Algae-based biochar also has different

characteristics both in physical and chemical terms. It has lower carbon content and surface area but has higher nitrogen and mineral content as well as various functional groups compared to lignocellulosic-based biochar, which makes it ideal for soil amendment and biosorbent activities [117, 118].

High temperatures are essential for breaking down biomass bonds and converting them into solid, liquid, and gaseous products. However, to maximize biochar yield, a comparatively lower temperature should be selected. Higher temperatures cause the thermal decomposition of biochar, leading to increased liquid product yield [103]. Further increasing the reaction temperature promotes additional cracking and secondary reactions, resulting in a higher yield of gaseous products at the expense of solid char [111].

The heating methodology has a significant impact on pyrolysis. Generally, the biomass feedstock is heated by conduction, convection, or radiation using energy from the combustion of fossil fuels or in some cases electricity generated by renewable sources. However, in these traditional methods, the heating rate of biomass particles is typically limited due to the low thermal conductivity of biomass. Microwave heating, on the other hand, involves the internal heating of biomass particles through the absorption of electromagnetic radiation [119]. In addition to the reduction of the heating cost [120], there are advantages like uniform heating of biomass particles [121], lower pyrolysis temperature [121], improved pore structure, and higher biochar quality [122]. Another key aspect that has a high influence on the products is the rate of heating; for example, fast heating rates produce rapid fragmentation of biomass and thereby increase the yield of volatiles [123]. At high heating rates, secondary pyrolysis reactions (like depolymerization of cellulose and hemicellulose) have a significant role in the reduction of the formation of biochar while enhancing the formation of the gaseous and liquid components [124]. On the other hand, a low heating rate prevents this kind of thermal cracking of the biomass, resulting in a greater biochar yield.

During the pyrolysis process on biomass, an inert environment is required at all times; i.e., carrier gas N_2, H_2, CO, CO_2, CH_4, and steam can be used [124]. As for the pressure, it was reported that higher pressures may enhance the biochar formation because the pyrolytic gases come into contact with the solid products more efficiently under elevated pressures

[125]. The carbon content of the biochar is also affected by the high pressure inside the reactor. The carbon concentration in the biochar is reported to increase under high pressure, which in turn increases the energy density of the biochar [126].

When the residence time of a pyrolysis process increases, many complex events occur, changing the composition of the product. However, as discussed above, residence time alone is not the only variable that affects the product. As a result, general principles for gas and liquid yield are difficult to obtain [127]. However, in general, long vapor residence time is needed for higher biochar production due to the fact that secondary processes such as cracking and repolymerization may occur when vapors remain in the reactor for an extended period of time [128]. In addition to the effect on the biochar yield, residence time promotes the development of micro and macro pores, which impact biochar quality [103]. Biomass has weak thermal conductivity, creating challenges for heat transfer during pyrolysis [111]. Consequently, particle size significantly impacts yields of bio-oil and biochar. Smaller particles are generally preferred when bio-oil is the desire product, because they heat up more uniformly than larger ones [123]. Insufficient heat transfer in larger particles results in lower liquid yields, whereas larger particles increase biochar formation [103].

Pyrolysis vapor is condensed into bio-oil during the fast pyrolysis process. However, the upgrading of the bio-oil is not thermally efficient since it requires evaporation and condensation simultaneously. The introduction of catalysts into the pyrolysis process before condensation, on the other hand, is more feasible because upgrading occurs directly after or at the same time as the pyrolysis [129]. Catalytic pyrolysis is one of the most promising methods to improve the quality of bio-oil by removing oxygenated compounds, increasing calorific value, decreasing viscosity, and enhancing stability [130]. Nevertheless, catalytic pyrolysis may have some disadvantages like the quick consumption of the catalyst, having a short life-cycle of the catalyst due to poisoning or deactivation, and the presence of the catalyst leading to an increase in the solid residue [131].

Co-pyrolysis, which has gained popularity as an optional strategy for obtaining high-grade products with or without using any catalyst, is a process in which two or more materials (like wood biomass and waste plastic) are utilized as feedstock. The key to the success of this technique lies primarily in the synergistic effect that results from the

interaction of various radicals formed from different materials during the process, which leads to a more stable pyrolysis oil [132]. Although co-pyrolysis is a cost-effective alternative process, it is necessary to adjust the types and proportions of feedstock to obtain a high-grade product. Since this technology uses several types of biomasses, an additional pre-treatment system may also be needed [133].

2.1.4. Lipid Production from Oleaginous Yeasts

Biodiesel production from lipid-based biomass is a sustainable solution to meet the ever-increasing fuel demand while inflicting little environmental impact due to the rapid bioenergy cycle. Despite the fact that biodiesel may be made from any edible plant-based oil or animal-based fat, these first-generation feedstocks are no longer favored due to the need for extensive agricultural areas, competition with the food market and high raw material cost [134]. Instead, biodiesel made from microalgae, cyanobacteria, or oleaginous yeasts is preferred more because these organisms can grow very quickly compared to other bioresources and they do not compete with food production [21].

Oleaginous yeasts can synthesize and accumulate lipids, where some species can accumulate up to 70% of their dry cell weight (DCW), although they represent a small proportion of the total yeast population [135]. Only 5% of these yeasts are capable of accumulating more than 25% lipids. Yeasts offer several advantages for lipid production over other microbial sources, including shorter duplication times (often less than 1 hour), less susceptibility to seasonal or climate conditions compared to plants, and easier scalability of cultures compared to microalgae [136].

Yarrowia, *Candida*, *Rhodotorula*, *Rhodosporidium*, *Cryptococcus*, *Trichosporon*, and *Lipomyces* are examples of oleaginous yeasts in general [137]. Among them, *Yarrowia lipolytica* is a promising cell factory for lipid production [138]. It is an aerobic yeast that can consume a wide variety of carbon sources, including various oils, fats and hydrocarbons [139]. In addition, *Yarrowia lipolytica* can utilize alternative carbon sources with or without metabolic engineering and/or adaptive evolution, especially from lignocellulosic biomass [140]. For instance, xylose [140, 141], olive mill wastewaters [142] and sugarcane bagasse [143] are some of the renewable feedstocks that are investigated in the literature. It has

enormous biotechnological potential for producing single-cell oils as well as citric acid and other organic acids [144]. Furthermore, oleaginous yeasts are good candidates for adaptive laboratory evolution or evolution-based strategies to increase lipid accumulation [145, 146]. As a result, the interest to use yeasts (especially *Yarrowia lipolytica*) to produce biodiesel has accelerated in recent years, as evidenced by the number of related publications presented in Figure 2.7.

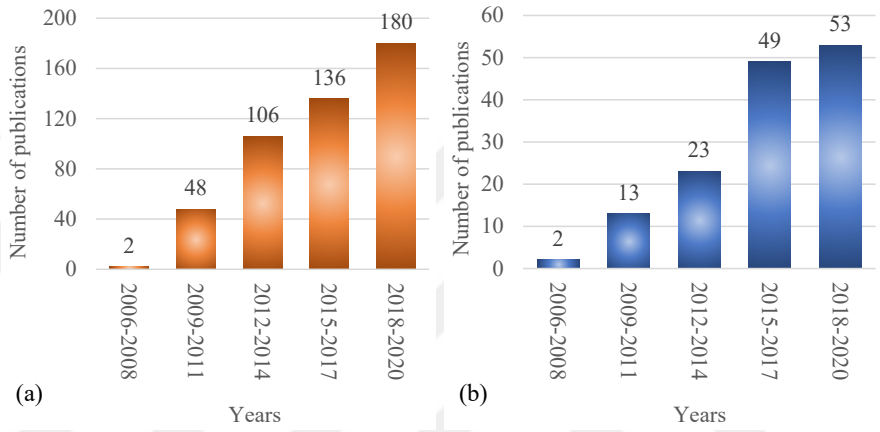

Figure 2.7. Number of publications through years in the "Web of Science Core Collection" database for: (a) "yeast" in the title and "biodiesel production" in the topic, (b) "*Yarrowia lipolytica*" in the title and "biodiesel production" in the topic; searched on 27 July 2021 [154].

There are numerous publications in the scientific literature reporting large numbers of experimental results aimed at optimizing biomass growth and lipid production of *Yarrowia lipolytica* for biodiesel production. For instance, there are studies in the literature that aims to optimize the cultivation conditions for high yields of biofuel production from those oleaginous yeasts by employing composite factorial design and response surface methodology [147-149] and also numerical models for predicting biomass production [150]. However, conventional methods make it difficult to determine a single best condition due to the complicated nature of the experimental and operating variables in these works. Instead, machine learning tools can assist in analyzing such kind of a database to detect the major patterns that are difficult to spot with the naked eyes and to determine the optimal process variables [151].

A good candidate oleaginous yeasts species for biodiesel production should ideally grow to high cell densities with a high lipid content and exhibit robust growth characteristics at low pH and a broad temperature range, facilitating industrial process development [152]. Several factors influence the growth and lipid synthesis of oleaginous yeasts, which can be categorized as; culture media (carbon and nitrogen source, carbon-to-nitrogen ratio (C/N), temperature, pH, etc.), cultivation mode (batch, fed-batch, repeated batch, continuous, etc.), genetic engineering, and other techniques as co-culturing [153].

2.1.5. Tri-reforming of Methane

The use of biomass as a source for biofuel production represents an economical and environmentally friendly alternative to traditional fossil fuels. Moreover, biogas produced from biomass is a good substitute for natural gas. The primary components of biogas are CO_2 and CH_4, with trace amounts of impurities whose ratios vary depending on the source. The most common methods for biogas production are the anaerobic digestion of sewage sludge or animal manure by microorganisms (containing 60–80% CH_4 and 20–40% CO_2) and the landfill gas process which is the decay of organic garbage in the absence of oxygen (containing 40–60% CH_4 and 40–60% CO_2) [155].

The good thing about biogas is that methane in the biogas can be further utilized to produce synthesis gas (syngas). Indeed, syngas is a valuable chemical intermediate that is used for producing valuable chemicals and fuels, such as methanol, dimethyl ether, ethylene, propylene, and etc. [156].

Syngas can be produced through processes such as steam reforming of methane (SRM), dry reforming of methane (DRM) and partial oxidation of methane (POM). Among these processes, SRM is highly endothermic and produces syngas with a high H_2/CO ratio of 3, which can be utilized for ammonia synthesis. On the other hand, DRM produces a low H_2/CO ratio around unity, which is more suitable for dimethyl ether production; however, the DRM process is constrained by rapid catalyst deactivation due to carbon deposition. Finally, POM has the advantage of being the sole exothermic reforming reaction producing syngas with an H_2/CO ratio of around 2, which is ideal for methanol production [155, 157, 158].

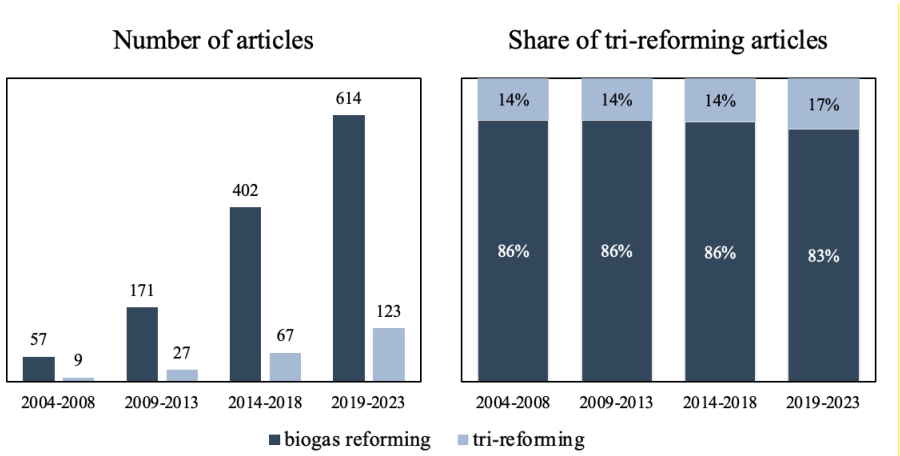

Figure 2.8. (a) Number of publications through years in the "Web of Science Core Collection" database, and (b) share of tri-reforming articles; searched on 25 May 2024.

Tri-reforming of methane (TRM) has emerged as an innovative method to produce syngas by integrating DRM, SRM and POM reactions into a single operation [158, 159]. This methodology combines the advantages of all three processes producing syngas with an H_2/CO molar ratio of around 2 (dominant reaction routes), which is suitable for producing methanol and Fischer-Tropsch fuels [160]. In addition, TRM proves to be an energy-efficient and viable alternative for syngas production by leveraging both endothermic and exothermic reactions within a single operation and utilizing waste flue gas and biogas without prior separation process of its components [156, 160]. The introduction of O_2, H_2O, and CO_2 alongside CH_4 into the reaction offers several advantages, as first introduced by Song and Pan [161]. Such that, the presence of H_2O and O_2 serve as oxidizing agents and mitigates coke formation by converting carbon deposits into CO and CO_2, thus enhancing catalyst stability [155, 160]. Figure 2.8 illustrates the growing interest in biogas reforming in general, as well as the share of TRM in this. The figure indicates that the number of published articles on both biogas reforming and tri-reforming has increased through the years. Furthermore, the number of TRM-related publications appears to have increased even more in recent years.

In TRM, depending on the feed compositions (i.e., CO_2, O_2 CH_4, H_2O), and reaction conditions, process can be mildly endothermic, thermoneutral or mildly exothermic [162]. Although TRM is a very innovative method for the production of syngas, it is also quite complex involving many variables to optimize. As a result, there are numerous experimental

and computational studies on the TRM process in the literature, with the goal of determining the optimal process design, development, and operations for syngas production, as well as understanding the underlying mechanism of biogas tri-reforming, coke formation, and equilibrium conversion. For instance, simulation studies for the TRM process are carried out [163], with membranes [164, 165], in fuel cells [166] coupled with the conversion of syngas to methanol [167, 168] and coupled with Fischer-Tropsch [169]. Also, kinetic modeling with thermodynamic analysis is carried out by some researchers [170-172]. The most common reaction that takes place can be given as follows:

$CH_4+H_2O \leftrightarrow CO+3H_2$	ΔH^o_{298K}=206.3 kJ mol^{-1}	Steam reforming	(2.1)
$CH_4+ 1/2\ O_2 \leftrightarrow CO+2H_2$	ΔH^o_{298K}=-36.0 kJ mol^{-1}	Partial oxidation	(2.2)
$CH_4+CO_2 \leftrightarrow 2CO+2H_2$	ΔH^o_{298K}=247.0 kJ mol^{-1}	Dry reforming	(2.3)
$CH_4+2O_2 \leftrightarrow CO_2+2H_2O$	ΔH^o_{298K}=-880.0 kJ mol^{-1}	Complete oxidation	(2.4)
$CH_4 \leftrightarrow C+2H_2$	ΔH^o_{298K}=75.0 kJ mol^{-1}	Methane cracking	(2.5)
$2CO \leftrightarrow C+CO_2$	ΔH^o_{298K}=-172.0 kJ mol^{-1}	Boudouard reaction	(2.6)
$CO+ H_2 \leftrightarrow C+H_2O$	ΔH^o_{298K}=-131.0 kJ mol^{-1}		(2.7)
$C+ O_2 \leftrightarrow CO_2$	ΔH^o_{298K}=-394.0 kJ mol^{-1}		(2.8)
$CO+ H_2O \leftrightarrow CO_2+H_2$	ΔH^o_{298K}=-43.0 kJ mol^{-1}	Water gas shift reaction	(2.9)

Machine learning (ML) may also play a role in determining the combination of variables that leads to the highest syngas yield. ML can be employed to analyze large datasets and deduce trends using algorithms like artificial neural networks (ANNs), support vector machines (SVMs), random forest (RF) regression, and decision trees (DTs), and these methodologies have also been applied in reforming processes in recent years. For instance, ML is performed in the DRM process to model; the deactivation of alumina-supported cobalt catalysts [173], CO_2 and CH_4 conversion over nickel catalysts [174], H_2/CO ratio on nickel catalysts [175], feed conversions and H_2/CO ratio of the product stream [176], syngas composition from bio-oil [177], CH_4 and CO_2 dissociation energies and barriers with bimetallic catalysts [178]. Also, SRM is modeled with ML; to classify conversion rates [179], predict syngas flow rate, product compositions, and steam temperature [180], CO_2 emission and thermal efficiency [181]. Water-gas shift reaction (WGSR) is modeled with

ML by Odabaşı *et al.* [182], autothermal reforming is modeled by Gul *et al.* [183] and partial oxidation is modeled by Ugwu *et al.* [184]. Bi-reforming is modeled by Deng *et al.* [185].

As it is explained above, TRM is a quite complex process with various reactions involved; such that, the produced syngas ratio from TRM and conversion of CH_4 and CO_2 depend not only on the feed composition but also on the operating conditions and the catalysts used [159]. As a result, there are numerous studies in the literature with so many experimental findings on tri-reforming of biogas to optimize the factors that lead to the desired syngas composition. However, due to the complexity of the experimental variables and the large size of the data, it is difficult to determine the best conditions for methane tri-reforming with the naked eye.

2.2. Machine Learning: Basics and Challenges

ML is a branch of artificial intelligence to creates programs that can automatically learn from the large datasets using statistics and some algorithms; it performs some functions like prediction, classification, clustering and association to analyze the data, recognize patterns, make predictions or extract rules [186]. Implementation of ML can be divided into four major steps: *data gathering*, *pre-processing*, *model development* and *evaluation and utilization of model*. In *data gathering* step, preferentially a large dataset correlating descriptors (input variables) and desired performance (or outcome) is created. Both quality and the quantity of the data are crucial at this stage; high quality is needed to avoid biases and incorrect conclusions while significantly large number of data (relative to number of descriptors) is needed for the high generalization ability of the model created. Both experimental and computational (like DFT created) data are used for ML models while the source of the data may be also different (like creating *in house*, collecting from literature, or extracting from databases) [187]. For example, Materials Project Database (MPD), as one of the recently created material databases, contains physical/chemical properties of more than 530000 nanoporous materials and 124000 inorganic compounds; it also provides tools to analyze the data [188]. However, such databases are mostly for relatively simple inorganic materials, and field specific databases for complex systems like biofuel productions are not that common.

In pre-processi*ng step,* the data are formatted to a machine-readable form while the missing information is completed if it is possible (otherwise incomplete data points are removed). The potential descriptors should be analyzed and their numbers should be reduced as much as possible (dimensionality reduction) because the smaller models (for a given dataset) are more robust [189]. This could be achieved by eliminating the descriptors that are not significant for the output or highly correlated with other variables (*feature selection)* as well as creating a new but fewer descriptors using techniques like principal component analysis (*feature extraction*).

Then, the *model development* step can be implemented by starting with selection of the algorithms to be used (highly dependent on the purpose of the research and characteristic of the data); Figure 2.9 indicates the major tasks and algorithms used in ML models. Regression is the most common task performed through algorithms like simple regression, artificial neural networks (and deep learning networks), support vector machines, regression trees, and gradient boosting trees [190]. Classification is another frequently implemented task (usually for categorical variables) through the algorithms such as decision trees (DT), support vector machines (SVM) and logistic regression (LogR) [191]. The common practice in model building is that the dataset is divided into two part as training (model building) and testing (verification of model performance), the training is performed using the first set together with a validation procedure (like k-fold cross-validation) to select the best model (i.e. to determine the model hyperparameters). Finally, the selected model is tested using the testing data for its ability to predict the data not seen during model development. In evaluation and utilization of model stage, the results are interpreted and utilized for various purposes such as making predictions, selection of materials or methods for a specific purpose, understanding the effects and interaction of descriptors and developing heuristics rules or guidelines for the future works.

There are also limitations in ML algorithms. Most important limitations of ML are that it relies on the past data and the model developed should be valid within the limits of the datasets; hence, it may not provide or suggest breakthrough solutions, which usually come from unexplored areas. However, it makes significant contributions to understanding the physical phenomena and guiding the experimental research into more promising areas. Another potential problem is that most of the ML algorithms are more accurate in deductions

involving the materials, methods or process conditions that have large number of repeat in the database; it is quite possible that some new alternatives may be ignored due to its very small number of appearance in the dataset (too small for generalization) or decided to be insignificant due to the initial (immature) poor performances even though they may have great potential for the future like many great innovations. There are also pitfalls that are not inherited in ML or its algorithms but they are related to implementation. For example, sometimes ML algorithms are not selected by considering the purpose of research or structure of the data as normally should be; instead, they are selected based on their popularity at that time without providing any real benefit. Another common problem in ML application is that the ML model developed may be actually over-fitted (especially if the available data is limited) because the criteria for over-fitting for some ML algorithms are as not well defined (or obvious to observe) as those in linear regression. Similarly, the use of highly correlated descriptors may not be always avoided even though it is problematic in some ML algorithms. Consequently, ML models should be used with care and supported by the other tools and observations.

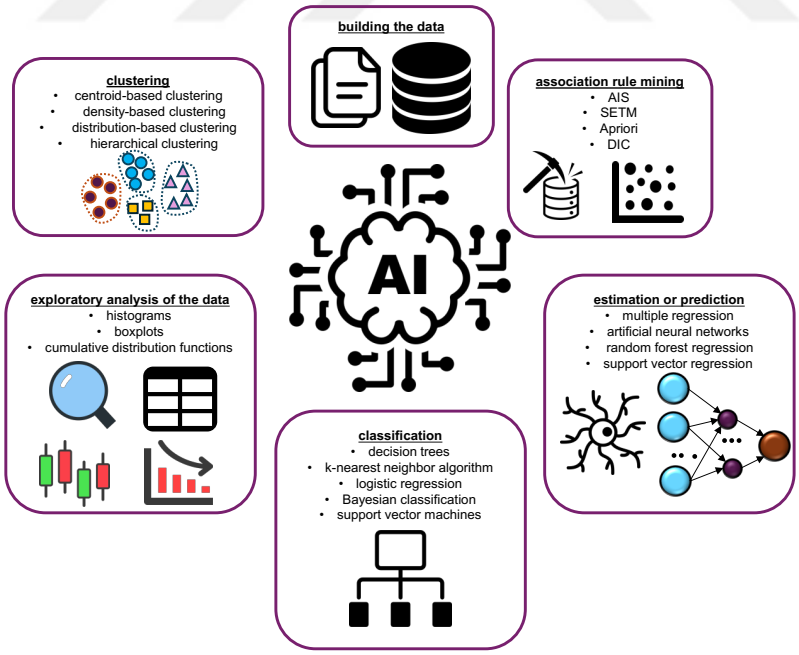

Figure 2.9. Conventional machine learning tasks and some popular algorithms.

Machine learning 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. As most of the other fields of scientific research, one of the biggest challenges for ML in biofuel production is to construct a dataset suitable for the knowledge to be extracted. ML relies on the data that represent the physical problem well; the information content of the data should be suitable for the problem in hand while the accuracy should be sufficiently high. In addition, the dataset should be sufficiently large for statistical significance, and typically very large datasets (as the related term of *big data* indicates) are required for complex systems with a large number of interdependent descriptors (*i.e.*, input variables). Building large datasets are relatively easy in *material* discovery and selection; desired properties or functions can be linked to more fundamental physical and chemical structures that can be determined experimentally (*e.g.*, spectroscopic characterizations) and computationally (*e.g.*, DFT simulations) even though their effectiveness can be questioned. Indeed, very large number of material related databases such as Material Project [192], OQMD [193], AFLOWLIB [194], Computational Material Repository [195], AiiDA [196] have been developed or extended in recent years with the consideration that these databases can be used to discover, select or analyze the functional materials (*e.g.*, catalysts, semiconductors, absorbers, membranes, *etc.*). While the numbers and sizes of such databases are growing, the accuracy of their data content have been increased as well thanks to the continuous development in experimental and computational technologies. Furthermore, these databases are also becoming easier to use with the available workflow management programs as AFLOW [194], Firework [197], Atomate [198], and pymatgen [199]. However, there are two major problems that should be considered. First, most of the current material databases contain information related to the general properties of the materials; hence, they are not suitable for field specific applications. The development of field related databases is also underway; though, they are relatively much smaller in their numbers and sizes due to the specificity of the data needed and smaller number of contributors for the construction of databases. Increasing efforts to develop new algorithms that can be used with smaller dataset may ease the problem [200]. However, the needs for the large datasets will always exist; it almost comes with the territory. Second, the qualities and performances of individual materials or components do not always reflect their performance in larger systems, where the interactions between the elements are also critical. The data for the system level performance are mostly available in the published papers, and only few groups utilize them in ML applications because not only the data extraction is tedious and time consuming but

also the dataset constructed from multiple sources contain significant level of noise and mismatches. Such tasks are becoming much easier with the recent improvements in ML such as *deep learning* for more effective knowledge extraction [97, 201], *transfer learning* to transfer knowledge from well-established data together with ML models to similar but less known cases [202] and physics informed machine learning to incorporate the physical laws and knowledge to data-driven models [203].

Explainable machine learning or artificial intelligence [204, 205], as a recently popularized trend in ML application may be also relevant to biofuel production research. As occasionally criticized, most of the ML algorithms end up with black box models that give little insight about the nature of descriptor-target variable relations. Explainable machine learning algorithms are aimed to overcome this weakness and maximize the learning from ML models. Although small datasets with high number of descriptors are a problem, there are also approaches that can be used for small datasets as 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) [189]. Meanwhile, new ML algorithms and approaches to be used for small datasets have been also investigated in recent years [200, 206, 207].

3. REVIEW AND ANALYSIS OF MACHINE LEARNING IN BIOFUELS

3.1. Microalgae Biodiesel

3.1.1. Bibliometric and Keyword Analysis of Microalgae Studies

Bibliometric evaluation of scientific literature is frequently 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 [208]. In order to have a comprehensive understanding of the current developments in the field of microalgae research, the publication trends were analyzed, a keyword search was carried out over the scientific literature. For this purpose, 7316 articles from Web of Science database were used with the search term *microalgae.* Author keywords with respect to publication years were used to find the trends in the microalgae research area; the duplicated and synonym keywords were combined. For a better comprehension, the related keywords were grouped into seven major categories: *general/concepts, cultivation, microalgae, products, conversion, modeling* and *metabolic*; *general/concept* category contains the keywords representing the basic concepts or issues in the fields. The results are summarized in Figure 3.1. To eliminate the fluctuations in years and to see the trends more clearly, data obtained from 2011 to 2022 were presented as a three-year moving average.

As clearly observable in Figure 3.1a, the number of papers increases with time (2022 bar contains data until November 4, 2022) while the frequency of *concept* and *conversion (*also *modeling* and *metabolic*) keywords, which are more likely to represents the recent trends, are increased as expected; the percent values represent the percent of articles containing a keyword related with that category in microalgae papers published in those years (hence, they do not have to be 100 in total). Figure 3.1b shows the change in the frequency of *cultivation* and *conversion* related keywords as the indicators of the trends in these major processes. *Two stage cultivation* appeared more frequently in recent years even though the total number is as not high as the single step processes yet. Although *hydrothermal liquefaction* is the most frequently appeared process together with lipid

extraction, which is mostly linked with transesterification, the attention to *pyrolysis* and *fermentation* processes seems to increase more in recent years. On the product side, the frequency of *FA/FAME/TAG* group does not change anymore even though they are still most frequently reported products (Figure 3.1c); the frequency of bioethanol, biohydrogen and bio-oil, on the other hand, steadily increased in recent years while some other value added products such as carotenoids and metabolites are also appeared more frequently.

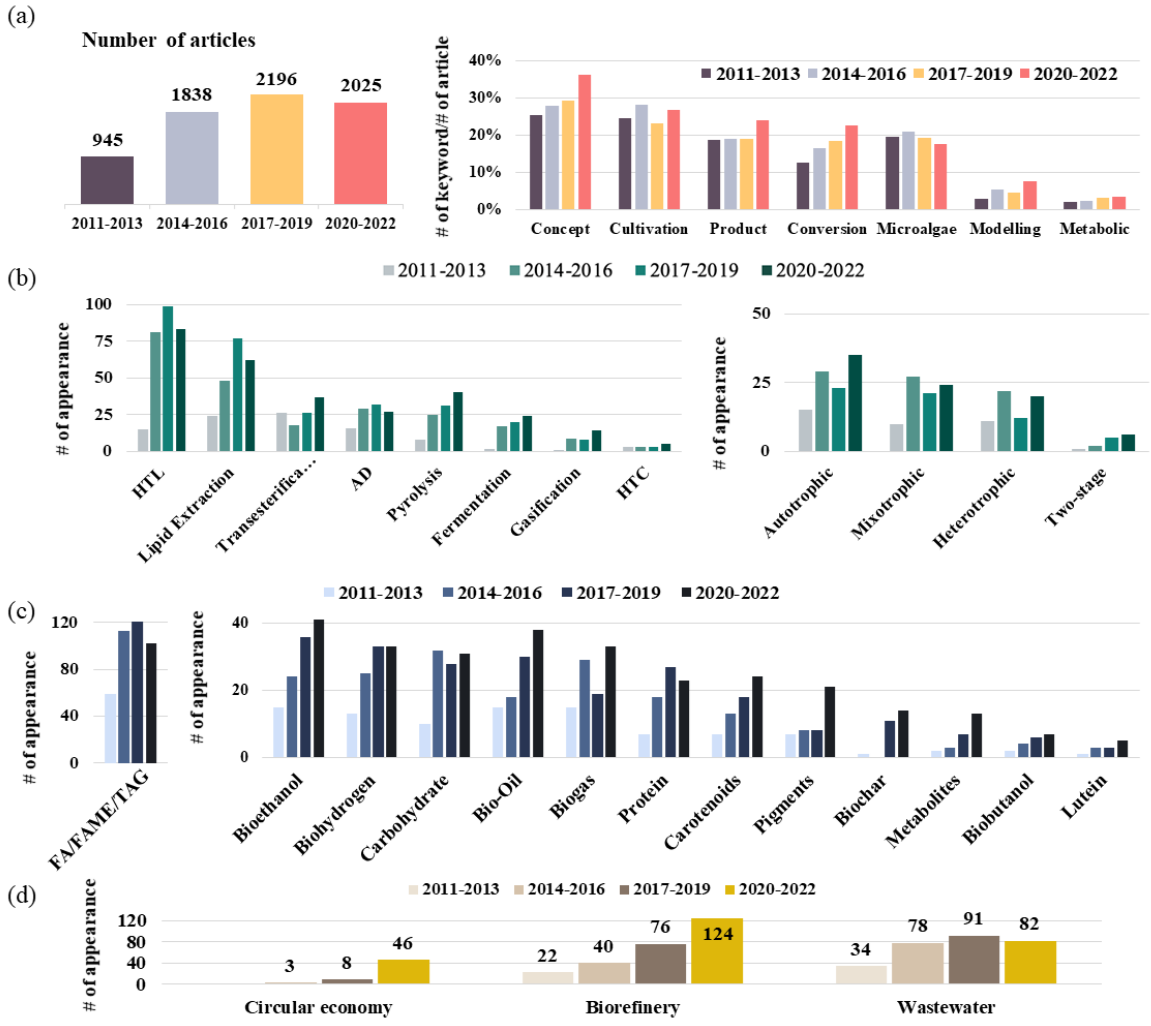

Figure 3.1. Change in frequency of keywords appeared in literature through the years [25].

In addition to those related with processes and products, some keywords are also worth to mention to see the general trends. For example, *wastewater, biorefinery* and *circular economy* keywords appear quite frequently in the author keywords; while the frequency of wastewater keyword seems to be declining even though the number of articles increases, the

popularity of *biorefinery* concept, which include biomass conversion, extraction, and separation of bioproducts and bioenergy [209], appears quite often, as shown in Figure 3.1d. *Circular economy* concept seems to be also in the center of microalgae studies in recent years [210]. Additionally, the appearance of *open pond* keyword is decreasing while *machine learning,* together with metabolic related keywords are increasing.

3.1.2. Machine Learning on Algal Biofuels

Kinetic modeling and data-driven machine learning (ML) are two basic approaches that have been used to model biological systems [8]. Traditionally, kinetic models (*e.g.*, Monod, the Droop, and the logistic models) are used to simulate the bio-production systems [211]. Modeling, optimization and state estimation of fed-batch systems have been successfully performed using kinetic models [212]. However, the predictive capacities of these models are usually low due to the reduction of a vast number of metabolic reactions into a few parameters and small sets of differential equations; any changes in the base conditions may have significant effects on model accuracy [12]. Data based models have been used as an alternative to kinetic based models in recent years because they can better predict a wide range of untested conditions than kinetic models because they use larger datasets that are valid for a wider spectrum of operating conditions and more variables for better description of the characteristics and behavior of process [12].

In recent years, some review papers were published on data-driven modeling of biofuel production [96, 213-215]. For instance, Ning *et al.* reviewed the performance of ML algorithms used on microalgae species identification and monitoring, and also summarized their advantages and disadvantages [215]. Wang *et al.*, on the other hand, presented a comprehensive review on ML application on bioenergy, microalgae cultivation, biofuel conversion processes without any particular focus on microalgae applications [96]. Ascher *et al.* focused on the application of ML on thermochemical conversion methods of various biomass including microalgae and waste [214], while Agbashlo *et al.* systematically reviewed biodiesel production from various biomass sources, their certification, and engine application [213]. Sundui *et al.* reviewed ML applications in biological wastewater treatment using algae-bacteria consortia [216]. These publications partially cover the microalgal biofuel production as well; however, does not give a complete picture by covering all steps

of algal biofuel production together with potential improvement areas and courses of action that can be taken.

In this section, research articles involving the analysis of microalgal biofuels using ML are reviewed in detail. For this purpose, the scientific literature was searched through academic databases (*e.g.*, Web of Science, Scopus, and Google Scholar) with keywords of *microalgae* together with *machine learning* and *data mining* (as well as algorithms like *artificial neural network)*, and 41 articles were identified to represent the subject well. Then, these articles were categorized with respect to the process involved as well as the ML tasks and algorithms applied. As expected, most of the studies have been published in last few years. The information about the nature of works reported in these publications are also presented in Figure 3.2. For instance, most of the publications were related with the cultivation step while microalgae classification was also frequently studied. The most common ML task performed was regression; neural networks (NN) is the most widely employed method with the contribution from the other techniques such as support vector machine (SVM) and extreme gradient boosting machine (XGB).

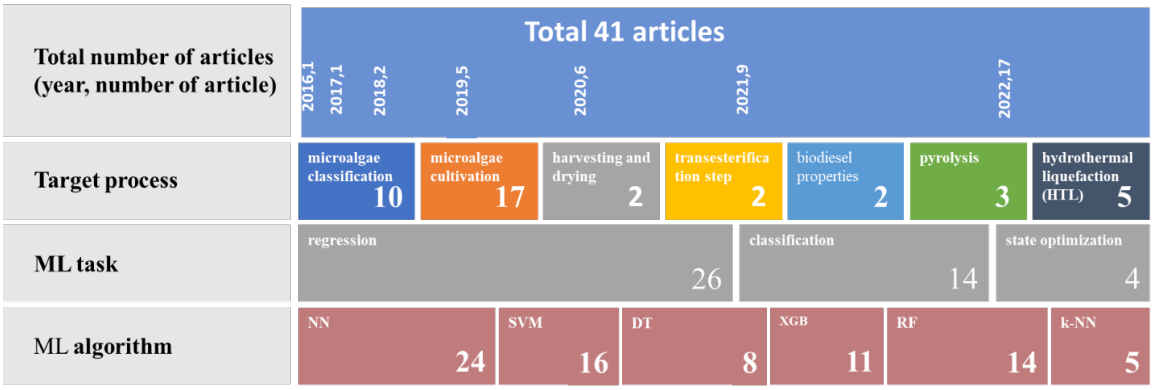

Figure 3.2. Summary of the machine learning articles in microalgae study [25].

The works published on the subject so far indicate that artificial neural networks (ANNs) worked well in most regression tasks performed in the fields, and it was used frequently to optimize the cultivation medium, which may be a good case to demonstrate the potential contribution of ML. In general, response surface methodology (RSM) is frequently (and effectively) used for this purpose. However, Condor *et al.*, Hossain *et al.*, and Srivastava *et al.* compared ANN with RSM, and all found that ANN performs better in given

tasks [217-219]; this observation was also stated by Desai *et al.* and attributed to the higher generalization capacity of ANN than RSM [220]. Apparently, ability of ANN in handling highly nonlinear systems (and flexibility in data structure compared to RSM, which require a rigid statistical experimental design) makes it a better candidate to perform regression tasks in cultivation.

For classification tasks, the main method of choice is decision tree (DT). In state optimization ANN, hybrid, long short-term memory (LSTM) and GP all used. Together with neural networks, tree-based methods such as XGB, random forest (RF), and DT are also frequently used. Zhang *et al.* stated that tree-based algorithms successfully used in classification and regression tasks of thermal conversion process in the biomass valorization domain [221].

Table 3.1, Table 3.2, and Table 3.3 summarize the details of all 41 ML articles on microalgae biofuels. It is worth to mention that there are other ML related studies in the microalgae research area even though they are not included in the tables (and the bibliometric analysis presented above) because they are not directly related to biofuels. For example, Panahi *et al.*, studied the transcriptomes of *Dunaliella salina* and *Dunaliella tertiolecta* species using machine-learning algorithms and RNA-seq meta-analysis [222]. Similarly, Kuo *et al.* reviewed big data analysis, and ML applications on the discovery of objective traits and genes in metabolic pathways [209]. Pozzobon *et al.* studied on dual staining procedure and introduced ML processing of flow cytometry readings to the microalgal biotechnology community [223], while Levasseur *et al.* studied the effects of light intensity, duty cycle, cycle duration on the microalgal performances using ML [11].

As briefly discussed above, there are tens of thousands of microalgae strains but only less than a hundred have been studied [57]. This is also true for ML applications, and most of the ML works related to microalgae strains are involved in classification. For example, Reimann *et al.* studied the binary classification of dead and living microalgae using seven ML algorithms, and found that the random forest classification achieved highest overall model accuracy with 77.7% [224]; a similar analysis was also performed by Wang *et al.* [225]. Other ML studies in this field had various aims such as characterization of culture composition [5], identification of microorganisms [226], classification of Cyanobacteria and

Chlorophyta microalgae groups [227], classification of monoalgal and mixed algal cultures [10], microalgae species distribution and biomass prediction [228], and estimation of the growth stage [229]. Details of the ML works on monitoring and identification are summarized in Table 3.1.

As presented in Figure 3.2, cultivation is the most studied step with ML. For example, Condor *et al.* studied the carbohydrate concentration in Chlorella vulgaris using ANN with light intensity, initial nitrogen concentration, and nitrogen starvation period as input variables. They found the R^2 value for prediction as 0.975, which was higher than that of the one obtained with the RSM method [217]. Similarly, Supriyanto *et al.* investigated the effect of cultivation conditions (i.e., initial algal concentration, harvesting period, hydraulic retention time, addition of sodium acetate, solar radiation, temperature, NO^{-3} concentration, and pH) on biomass concentration using ANN regression model with 553 data points, and found that the results were quite satisfactory [230].

The other ML studies related to cultivation step can be summarized as biomass prediction [8, 21, 230-234], carbohydrate prediction [217], carbon bio-fixation rate prediction [210], chlorophyll-a concentration prediction [235], effect of toxicity on growth and lipid production [236], growth rate prediction [11, 237], light distribution pattern prediction [11], lipid content prediction [21], lutein content prediction [8, 12, 212], N and P removal rate prediction [218, 233, 234] and nitrate concentration prediction [8]. Relatively higher frequency of ML works on cultivation can be attributed to the significance of carbohydrate, lipid, and lutein content of microalgae as they stand as the precursors of biofuels and value added biochemical. Carbohydrate is the only substrate required for the production of bioethanol, biobutanol, and biohydrogen from algae biomass [237]. Lipid is the key molecule for biodiesel production while the lutein is a primary carotenoid, which has received great attention in several industries, is a high value product [212]. Carbon bio-fixation rate, chlorophyll-a concentration and cell growth rate are also indicators of biomass production and their accurate assessment are important for microalgae biotechnology and environmental monitoring [228].

The preferable microalgae strains should have high CO_2 utilization rate, low nutrient requirements, high lipid and carbohydrate productivity together with high value added

biochemicals [210]. Efficient prediction of light distribution patterns, toxicity level of the cultivation medium, N and P removal rate and nitrate concentration are also important considering that insufficient irradiation and poor cultivation medium are the factors decreasing the product yield and therefore hindering the commercialization of algal biofuel [237]. The cultivation related ML works are summarized in Table 3.2.

Harvesting and drying are energy-intensive yet necessary part of algal biofuel production; hence they may limit the commercialization potential of microalgae biofuel [238, 239]. Several ML works involving these steps were also reported (Table 3.2); for example, Ching *et al.* investigated vacuum drying process by modeling change in moisture content (%) with three ML methods: ANN, SVM, and XGB, as XGB yielded the lowest RMSE. Temperature, pressure, and delta moisture values of 19,515 data is used for model construction [240]. Similarly, Suparmaniam *et al.* modeled the bio-flocculation efficiency as a microalgae harvesting technology using ANN algorithm on *Chlorella vulgaris* [239].

Various methods can be used for biofuel production from microalgae while these methods can be also utilized for the other biomass sources. Hence, discussions in this chapter are limited with the ML applications in which at least some of the data were from algal biomass. The most commonly employed biofuel production route is the lipid extraction followed by transesterification even though some new technologies are getting significant attention in recent years. As an example, Srivastava *et al.* investigated oil conversion to FAME by ANN with temperature, time and oil to methanol molar ratio as inputs [219] while Sultana *et al.* modeled biodiesel yields from microalgae oil using ANN and SVM models yielded RMSE values of 0.0614 and 0.0408 respectively [241]. On the other hand, Salam and Verma modeled the important parameters of biodiesel for its commercial use in engines; the properties included performance, emission, and combustion behaviors such as specific fuel combustion, brake thermal efficiency, indicated efficiency, and scavenging efficiency [242]. Suvarna *et al.* also modeled cetane number, cold filter plugging point and oxidation stability using biodiesel samples [243]. Representative examples of ML analysis for transesterification to biodiesel and the property estimation of products are presented in Table 3.3.

Table 3.1. Summary of machine learning on microalgae monitoring and identification.

(superscripts refer to a: train, b: test, c: validation, d: prediction, e: MRE)

Machine Learning Task	Aim / Target Variable	ML method	Feedstocks	Data size	Descriptors (# of descriptors)	Performance Measures	Reference
classification	characterization of culture composition	ANN	*Scenedesmus almeriensis*, *Chlorella vulgaris*	289,705	(32) sample images.	training error did not exceed 4%, validation error 6.5%.	[5]
classification	identification of microorganisms	LogReg	microorganisms	600 images, 32,779 instances	-	accuracy= 24.9%	[226]
		k-NN				accuracy= 96.1%	
		SVM				accuracy= 97.2%	
		ANN				accuracy= 31.2%	
		QDA				accuracy= 26.8%	
classification	binary classification of living and dead microalgae	NB	*Chlorella vulgaris*	-	(14) 5 morphological and 9 fluorescence features.	accuracy= 68.8%	[224]
		k-NN				accuracy= 74.3%	
		QDA				accuracy= 74.3%	
		DT				accuracy= 77.4%	
		AB				accuracy= 77.4%	
		RF				accuracy= 77.7%	
		ANN				accuracy= 75.9%	
classification	identification of living and dead microalgae	DT	*Prorocentrum lima*	1000	(10) digital holography microscopy.	accuracy= 94.2%	[224]
		LogReg				accuracy= 94.5%	
		SVM				accuracy= 94.8%	
		k-NN				accuracy= 89.7%	
classification	classification of Cyanobacteria and Chlorophyta groups	CNN	Cyanobacteria and Chlorophyta	472	inverted microscope images.	accuracy= 98.81%	[227]
		CNN-SVM				accuracy= 99.66%	
classification	classification of microalgae	CNN	microalgae	29,000	low resolution images acquired by a particle analyzer.	accuracy= 88.59%	[244]
classification	microalgae species distribution	CNN	microalgae monocultures	+100,000	optical properties of different microalgae groups.	accuracy= 95%[b]	[228]
			microalgae pairwise mixtures			accuracy= 100%[b]	
regression	biomass prediction		microalgae monocultures			MAE= 22%[b]	
			microalgae pairwise mixtures			MAE= 24%[b]	
classification	classification of microalgae	PCA-SVM	*Phaeocystis*, *Chlamydomonas* and *Chaetoceros*	180 (60x3)	hyperspectral imaging.	accuracy= 94.4%	[229]
regression	estimation of growth stage	RF				MAE= 0.781; R^2= 0.98; accuracy= 98.1%	
regression	description of monoalgal and mixed algal cultures	ANN	*Nostoc* sp., *Scenedesmus almeriensis*, *Spirulina platensis* and *Chlorella vulgaris*	550 samples	(31) Light absorption measurements.	R^2= 0.951 to 0.970	[10]

Table 3.2. Summary of machine learning on microalgae cultivation and harvesting & drying processes.

(superscripts refer to a: train, b: test, c: validation, d: prediction, e: MRE)

Machine Learning Task	Aim / Target Variable	ML method	Feedstocks	Data size	Descriptors (# of descriptors)	Performance Measures	Reference
Microalgae cultivation							
classification	effect of toxicity on algal growth	GBDT	*Phaeodactylum tricornutum*	165 datasets	(7) microplastics concentrations, polymer sizes, algal cell sizes, algal species, ratio of polymer size/cell size, polymer species, and exposure time.	false negativity and false positivity were 0.1 and 0.19	[236]
classification	biomass productivity	DT	microalgae	4356	(17)	accuracy= 89.3%[a], 88.5%[b]	[21]
	lipid content			2022		accuracy= 81.5%[a], 80.3%[b]	
classification	nitrogen removal efficiency	DT	Chlorodendrophyceae, Chlorophyceae, Consortium, Cyanophycean, Eustigmatophyceae, Trebouxiophyceae Xanthophyceae	231	(16) microalgae type, pre-cultivation stage deciding factors, nutrient content of wastewater, and operating variables.	accuracy= 76.08%	[233]
	phosphorus removal efficiency			229		accuracy= 77.7%	
	biomass production			221		accuracy= 84.09%	
classification	nitrogen removal efficiency	DT	Trebouxiophyceae	239	(13) pre-cultivation stage deciding factors, operating variables, initial nutrient concentration.	accuracy= 72.34%	[234]
	phosphorus removal efficiency			239		accuracy= 72.5%	
	biomass production			214		accuracy= 78.57%	
	nitrogen removal efficiency		Chlorophyceae	200		accuracy= 73.91%	
	phosphorus removal efficiency			204		accuracy= 80%	
	biomass production			206		accuracy= 73.17%	
regression	light distribution pattern	SVM	microalgae	-	-	R^2= 0.993	[237]
	growth rate			-	-	R^2= 0.992	
regression	relative variation of the growth rate	RF	green microalgae	761	incident light intensity, ratio between the light and the dark period of the cycle, and total cycle time.	around 90% of the maximum explained variance	[11]
regression	early prediction of final biomass yield	RidReg	*Spirulina platensis*	42	OD, pH.	MAE= 0.144[e] g/L; R^2= 0.754	[231]
regression	biomass concentration	GP	*Desmodesmus* sp. F51	-	incident light intensity and nitrate inflow rate.	GPs comparable to ANNs for long-term dynamic bioprocess modeling	[8]
	nitrate concentration						
	lutein concentration						
regression	carbohydrate content	RSM	*Chlorella vulgaris* FSP-E	20	(3) light intensity, initial nitrogen concentration, and nitrogen starvation duration.	R^2= 0.9873 (0.9329[d])	[217]
		ANN		27		R^2= 0.9959 (0.9752[b])	

Table 3.2. Summary of machine learning on microalgae cultivation and harvesting & drying processes (cont.).

(superscripts refer to a: train, b: test, c: validation, d: prediction, e: MRE)

Machine Learning Task	Aim / Target Variable	ML method	Feedstocks	Data size	Descriptors (# of descriptors)	Performance Measures	Reference
Microalgae cultivation							
regression	carbohydrate content	ANN	mixed consortium	108	(18)	RMSE= 0.0655; R^2= 0.8403	[245]
		CNN				RMSE= 0.0529; R^2= 0.8966	
		LSTM				RMSE= 0.06; R^2= 0.8648	
		k-NN				RMSE= 0.0921; R^2= 0.6831	
		RF				RMSE= 0.0678; R^2= 0.8286	
regression	N removal (%)	RSM	*Chlorella kessleri*	-	(3) temperature, light-dark cycle, and nitrate-to-phosphate ratio.	RMSE= 14.5112; MAE= 11.5189; R^2= 0.8653	[218]
		ANN				RMSE= 1.1881; MAE= 0.7869; R^2= 0.9994	
		SVM				RMSE= 0.7819; MAE= 0.5079; R^2= 0.9996	
	P removal (%)	RSM				RMSE= 10.88967; MAE= 9.80447; R^2= 0.8293	
		ANN				RMSE= 7.847; MAE= 4.8057; R^2= 0.9122	
		SVM				RMSE= 5.35345; MAE= 1.8936; R^2= 0.9592	
regression	chlorophyll-a concentration	PSO-SVM	*Spirulina platensis*	2340	(8) pH, optical density, oxygen concentration, NO^-_3 concentration, PO_4 concentration, salinity, water temperature and irradiance.	R^2= 0.99	[235]
state / dynamic optimization	fed-batch lutein production	hybrid	*Desmodesmus* sp. F51	-	-	final lutein yields of 5.0 mg L^{-1}.	[12]
state / dynamic optimization	carbon biofixation rate	GP	*Chlorella vulgaris* (ISC-23)	83	(6) temperature, light intensity, carbon concentration, aeration rate, pH, and time.	RMSE= 0.2372[a], 0.0820[b]; MAE= 0.0814[a], 0.0542[b]; R^2= 0.8791[a], 0.9393[b]	[210]
state / dynamic optimization	dry-weight yield	LSTM	*Spirulina* sp. HH	20-day measurement	light intensity.	17% higher than traditional cultivation method, 10% higher than threshold-based method.	[232]
Harvesting and drying							
regression	delta moisture (%)	ANN	*Chlorococcum infusionum*	19,515 generated data	(3) temperature, pressure, and delta moisture.	RMSE= 0.0437	[240]
		SVM				RMSE= 0.0500	
		XGB				RMSE= 0.0308	
regression	flocculation efficiency	ANN	*Chlorella vulgaris*	~195	(2) time and flocculation parameters.	R^2= ~0.98	[239]

Table 3.3. Summary of machine learning on transesterification, biodiesel properties, pyrolysis and hydrothermal liquefaction.

(superscripts refer to a: train, b: test, c: validation, d: prediction, e: MRE)

Machine Learning Task	Aim / Target Variable	ML method	Feedstocks	Data size	Descriptors (# of descriptors)	Performance Measures	Reference
Transesterification							
regression	% conversion of oil to FAME	RSM	*Chlorella* CG12	20	(3) temperature, time, and oil to methanol ratio.	RMSE= 1.53; R^2=0.97	[219]
		ANN				RMSE= 0.92; R^2=0.99	
regression	biodiesel yield	ANN SVM	microalgae oil	30	(4) catalyst dose, oil to methanol molar ratio, temperature, and time.	RMSE= 0.0614: MAE= 0.0328; R^2= 0.965[d] RMSE= 0.0408; MAE= 0.0396; R^2= 0.984[d]	[241]
Biodiesel properties							
regression	performance, emission, and combustion behaviors	ANN	*Spirulina* sp.	80	(3) loading levels, blending ratios, fuel injection pressures yield.	R^2= 0.9608	[242]
regression	cetane number	MLR	biodiesel	202	(10) C12:0, C14:0, C16:1, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C22:1.	R^2= 0.52[b]	[243]
		SVM				R^2= 0.77[b]	
		RF				R^2= 0.88[b]	
		XGB				R^2= 0.89[b]	
	cold filter plugging point	MLR		267	(11) C12:0, C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C22:1.	R^2= 0.77[b]	
		SVM				R^2= 0.75[b]	
		RF				R^2= 0.91[b]	
		XGB				R^2= 0.87[b]	
classification	oxidation stability	LogReg		227	(7) C14:0, C16:0, C18:1, C18:2, C18:3, C20:1, C22:1.	accuracy= 89%	
		k-NN				accuracy= 84%	
		SVM				accuracy= 93%	
		DT				accuracy= 91%	
Pyrolysis							
regression	weight loss	ANN	*Chlorella vulgaris*	~5000	(2) heating rate, temperature.	R^2= ~0.8	[246]
regression	biochar yield	XGB	microalgae	64	(5) ratios of elementary compositions (H/C, N/C), ash content, pyrolysis temperature, and residence time.	R^2= 0.9974[a], 0.8440[b]	[247]
regression	pyrolytic activation energies	RF	algae/biodegradable waste/cellulose/ fossil/lignocellulose	281 datasets	(10) type and properties of raw materials, applied model-free method.	RMSE= 8.4668; R^2= 0.9964	[248]

Table 3.3. Summary of machine learning on transesterification, biodiesel properties, pyrolysis and hydrothermal liquefaction (cont.).

(superscripts refer to a: train, b: test, c: validation, d: prediction, e: MRE)

Machine Learning Task	Aim / Target Variable	ML method	Feedstocks	Data size	Descriptors (# of descriptors)	Performance Measures	Reference
Hydrothermal liquefaction (HTL)							
regression	bio-oil yield	GBR	microalgae	310	(15) elemental composition (C, H, O, N and S), the atomic ratio of elements (H/C, O/C, and N/C), the biochemical composition of algae (protein, lipid, carbohydrate, and ash), and HTL operating conditions (temperature, residence time, and solid content of algae).	RMSE= 4.69[b]; MAE= 3.56[b]; R^2= 0.90[b]	[221]
	oxygen content					RMSE= 1.68[b]; MAE= 1.23[b]; R^2= 0.90[b]	
	nitrogen content					RMSE= 0.68[b]; MAE= 0.53[b]; R^2= 0.90[b]	
	bio-oil yield	RF				RMSE= 5.72[b]; MAE= 4.34[b]; R^2= 0.87[b]	
	oxygen content					RMSE= 1.91[b]; MAE= 1.43[b]; R^2= 0.87[b]	
	nitrogen content					RMSE= 0.74[b]; MAE= 0.56[b]; R^2= 0.88[b]	
regression	bio-oil yield	DT	biomass (*e.g.*, algae, sludge, manure, and food waste)	448	(18) elemental composition, process conditions, solvent characteristics.	RMSE= 2.89[a], 8.07[b]; R^2= 0.97[a], 0.82[b]	[249]
		RF				RMSE= 3.66[a], 6.17[b]; R^2= 0.95[a], 0.89[b]	
		GBR				RMSE= 3.74[a], 6.14[b]; R^2= 0.95[a], 0.90[b]	
	multi-task; bio-oil yield, nitrogen content in bio-oil, and energy recovery from biomass	RF		229	(12) elemental and biochemical composition, process conditions.	RMSE= 3.07[a], 4.69[b]; R^2= 0.92[a], 0.80[b]	
regression	biocrude yield	RF	*Chlorella* sp.	-	(6) biochemical composition (*i.e.*, protein, lipid, and carbohydrates), HTL process conditions (*i.e.*, temperature, reaction time, and initial solid concentration).	RMSE= 6.29[b]; R^2= 0.81[b]	[250]
	biocrude energy yield					RMSE= 2.14[b]; R^2= 0.84[b]	
regression	biocrude yield	XGB	wet biomass (*i.e.,* macro and microalgae, food waste, animal manure, sewage sludge, and wastewater) and waste	325	(17) feedstock characteristics (biological and elemental properties) and operating conditions.	RMSE= 5.4137; R^2= 0.8711	[251]
		RF				RMSE= 6.0073; R^2= 0.8425	
		KRR				RMSE= 6.3053; R^2= 0.8201	
		SVM				RMSE= 5.9615; R^2= 0.8457	
	higher heating value	XGB				RMSE= 1.9846; R^2= 0.8194	
		RF				RMSE= 2.2857; R^2= 0.7353	
		KRR				RMSE= 2.5142; R^2= 0.6567	
		SVM				RMSE= 2.4259; R^2= 0.7134	
regression	predictions of biocrude yields	SVM	algae, lignocellulosic biomass, and model compounds	570	(18) feedstock type, biochemical composition, solids loading, reaction temperature, reaction time, heating rate, organic biocrude extraction solvent, reactor type, reactor size, and yields of char, gas, and aqueous phase.	RMSE= 9.4[b]	[252]
		DT				RMSE= 10.24[b]	
		ANN				RMSE= 9.02[b]	
		RF				RMSE= 8.34[b]	
		XGB				RMSE= 8.57[b]	

Microalgae biomass produces important products such as fatty acids, carbohydrates, and pigments, that can be directly used in the production of high value bioproducts [253]; however, a large amount of algal residue also creates significant waste [254]. Hence, the thermochemical routes such as pyrolysis, hydrothermal liquefaction, and carbonization are increasingly popularized in recent years for utilizing algal residues [247]; this is also true for ML applications (Table 3.3). For instance, Bong *et al.* modeled weight loss in pyrolysis [246], while Pathy *et al.* examined the biochar production yield [247]; in another work, Wang *et al.* evaluated the pyrolytic activation energy using ML [248].

Transesterification is an easy and established process to produce bio-oil from microalgae; however, it requires the energy intensive drying step to achieve high bio-oil yield. Some methods, such as hydrothermal liquefaction (HTL), on the other hand, do not need drying step [255]; hence, they may have higher potential. Indeed, several ML works also performed over HTL technology as well (Table 3.3). For example, Zhang *et al.* modeled bio-oil yield, and oxygen and nitrogen content in bio-oil [221], while Li *et al.* predicted bio-oil yield together with nitrogen content in bio-oil (%) and energy recovery from biomass (%) in a multitask modeling approach using random forest regression [249]. Katongtong *et al.* [251] and Cheng *et al.* [252] performed ML analysis for the prediction of biocrude yield from hydrothermal liquefaction.

The lack of sufficient data seems to be a more serious problem for ML applications in algal biofuel production than many other fields. First of all, algal biofuel production is not a single process; it is collection of steps that are dissimilar in their goals, parameters and challenges each requiring a different dataset with different descriptors. For example, selection of algae, cultivation, harvesting and converting the biomass contents into more valuable products, including biofuel require completely different technologies. Luckily, some steps overlap with other technologies; for example, cultivation should have some similarities with the cultivation of other biological agents while the parameters and problems in transesterification of algal lipid to biodiesels should not be much different than those applicable to the lipids derived from other sources. The experience in similar fields always eases the implementation of ML in algal research (and vice versa); *the transfer learning*, which refers to the use of knowledge gained from a specific area to analyze/model another

similar area [202] may be applicable as well. However, significant amount of data specific to each process in algal value chain will be always required; this need is especially important for the algal strain selection and cultivation.

As summarized above, the biomass content, potentials to produce specific value-added products and suitability to wastewater to be used may differ significantly for different microalgae strains; hence, the selection and optimization of strain are among the most critical tasks for successful commercialization of algal biofuels. In addition to the diversity of microalgae strains in nature, the genetically modified algal strains have also proven to enhance the photosynthetic efficiency and improve the lipid content of the microalgae increasing the number of strain alternatives further [256]. It should be also noted that the characteristic of stress conditions will be highly dependent on location (due to the locality of wastewater) and time (due to the changes in wastewater as the results of change in habits and regulations). Also, there is no compressive work on ML screening the large number of potential strains for a specific application so far as done in various other fields, and the main reason for this appears to be the lack of data for sufficiently large number of strains. Indeed, the published ML articles in this nature are limited to the selection of the best alternative among few already known strains for a specific performance measure in a specific set of cultivation condition. The large datasets, however, may have enormous impact on the potential contribution of ML to commercialization of the algal biofuels; it will not only improve the biofuel production efficiency, but also help to control the production of additional value-added chemicals, which are highly dependent on the algal strain used. Furthermore, the understanding of structure (including genomic information) and activity relationship, will ease the genetic modification of microalgae for the production of specific value-added products.

Also, ML applications similar to algal strain selection problem can benefit from the explainable ML significantly. Differences in the structures and properties of the strains determine their effectiveness to produce lipids and other desirable products through some metabolic pathways; more information deduced for these pathways and intermediate compounds will make ML more valuable for future works, especially those involving genetic modification.

Cultivation is another step that may be improved significantly using ML tools. Indeed, most of ML works on algal biofuel production so far have been aimed to analyze and optimize the variables in cultivation step process (Figure 3.2). This is an expected result because, regardless of cultivation type (open or closed system), it is difficult to determine the optimal conditions for biomass growth or lipid production using conventional approaches; instead, ML tools can be quite useful in detecting major patterns that are difficult to see with the naked eye, as well as determining the ideal combination of process variables for microalgal biomass production [21]. Clarity of target variables (e.g., biomass growth or lipid production) and descriptors (i.e., clearly defined medium compositions and operational variables) as well as vast experience with ML analysis in cultivation of various biological agents also makes this field highly suitable for such analysis. However, the nonstandard and noisy nature of the process create a big challenge for the full utilization of ML models in the field. First of all, the composition of the cultivation is inherently nonstandard; wastewaters from different sources have different nutrition and contamination while there may be also differences with respect to time as we briefly discussed above. This significantly reduces the generalizability of ML models and limit learning from the data. Secondly, solar irradiation is quite nonstandard even for the works performed in laboratories under solar simulators; the type and brand of light source, distance and angle of irradiation, material of transparent bioreactor and the composition of medium will change the amount of light received by microalgae. Needless to say, the situation is much worse for the tests performed under actual solar radiation due to the continuously changing climatic conditions. We can also add the evaporation of water and other noise factors in open pond cultivation. In one hand, standardizing experimental conditions can be only achieved in laboratory scale, and definitely creates a scale up problem due to uncounted factors in actual conditions. On the other hand, the experimental data generated under actual conditions will have significant noises, which is not desirable for ML models, and unique variables (e.g., specific metals or chemicals that exist only in certain wastes), which is much worse than noise for ML applications. Nevertheless, the development and use of standard testing and reporting procedure for microalgal cultivation, as in other technologies like solar cells [257] seems to be the best choice to be able to extract generalizable knowledge despite of the disadvantages mentioned above; needless to say this will also require close collaboration among the researchers in the field. This will not only improve the effectiveness of ML analysis in algal

cultivation but also improve the knowledge accumulation via systematic comparison of the results under similar conditions.

3.2. Lignocellulosic Bioethanol

3.2.1. Bibliometric and Keyword Analysis of Lignocellulosic Bioethanol Studies

To understand the trends in the "lignocellulosic biofuel" research field, bibliometric evaluation is carried out by analyzing the publication trends and "author keywords" of the articles in the literature. For this purpose, Web of Science database is used with the search term *lignocellulosic biofuel*, and a bibliometric study is carried out with a total of 6853 publications.

Research interest in the field is foremost assessed by the number of articles published each year. It is found that articles related to lignocellulosic biofuel are increasing year by year as expected (Figure 3.3a). However, this trend is common in different research fields as the total number of SCI-indexed publications is also increasing [208]. To discover the assistance of machine learning in the field, another search is conducted with the term; lignocellulosic machine learning, and compared with lignocellulosic biofuel, as shown in Figure 3.3a. It is observed that machine learning inclusion in the field is not only increasing in number but also expanding its relative share in total publications. "Author keywords" are extracted from the publications and categorized with respect to the type of biofuel, type of feedstock, and conversion method to understand the trends in the area. A data cleaning step is conducted by combining the keywords which are duplicated and synonyms. Also, four-year moving averages are taken for analysis to eliminate any fluctuations in years. The result of categorized keyword distribution with respect to four-year averages is presented in Figure 3.3 to uncover shifts in the research trend in the field.

Second-generation biofuels can be produced using a variety of 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 [88]. 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 3.3b, *bioethanol* is the most studied lignocellulosic biofuel. It is also observable that *biochar* and *biogas* are gaining more attention, as almost half of the related papers in these fields are published in the last 4 years. As shown in Figure 3.3b, the trend of the conversion methods is also in agreement 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, *anaerobic digestion*, *hydrothermal liquefaction*, and *hydrothermal carbonization* are increasing their individual share in keywords each year.

The type of lignocellulosic feedstock utilized for biofuel production is very critical for efficient and economic conversion. Lignocellulosic biomass can be categorized as agricultural and forest residues, forestry products, dedicated energy crops, municipal solids, and industrial waste [91]. The most commonly studied feedstocks in literature are categorized and shown in Figure 3.3c. 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 is increasing every year.

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 appeared keyword indicating its recent popularity (Figure 3.3d). *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 got more attention than *dilute-acid pretreatment*. The increase in the research interest in *deep eutectic solvents* (DESs) is also worth mentioning as their frequency almost doubled in the last four years. The most commonly studied microorganisms are also presented in Figure 3.3d; 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 *Saccharomyces cerevisiae* is the choice for fermentation throughout the years.

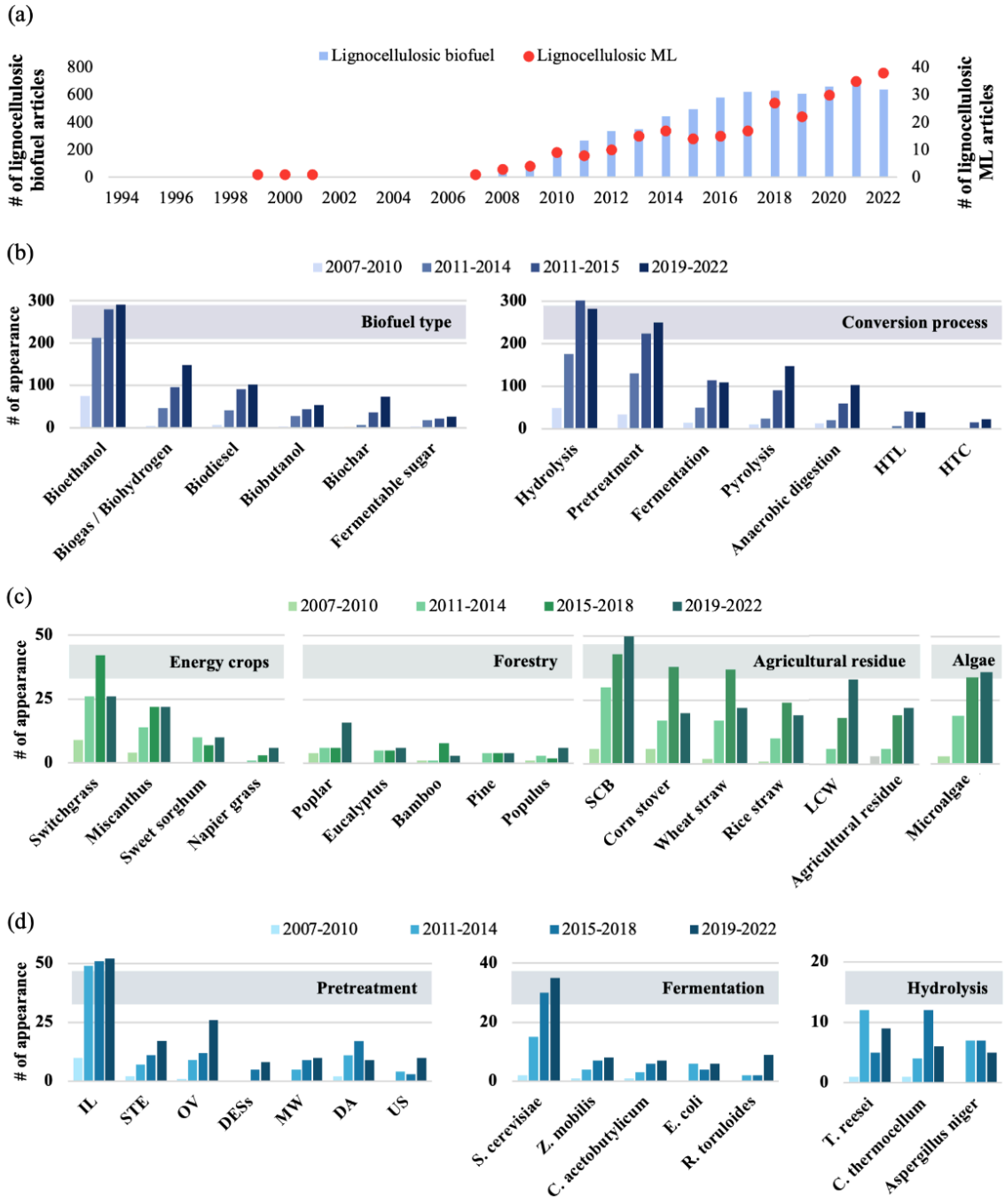

Figure 3.3. Publications trends in lignocellulosic biofuel through years: (a) total and ML-related number of lignocellulosic biofuel articles; keyword distribution of (b) biofuel and conversion processes; (c) biomass source; and (d) pretreatment methods, and microorganisms used for fermentation and hydrolysis [63].

3.2.2. Machine Learning on Lignocellulosic Ethanol Production via Fermentation Route

Due to the complexity of the process variables associated with the conversion of lignocellulosic biomass to ethanol, ML can help to determine the optimal experimental conditions leading to the highest conversion with the most feasible combination of variables. Indeed, the use of machine learning has grown significantly over the past few years in various areas including energy and fuels. There are also very successful applications of machine learning in biofuel research including some reviews covering the entire field [96] or specific studies such as ANN applications [258, 259], or reviews of ML for gaseous biofuels [88]. Applications of ML for lignocellulosic biomass have also been reviewed by a variety of research groups [260, 261]. Reviews concentrating on waste treatment [262], pyrolysis [263] and pretreatment of lignocellulosic biomass [76] are some of those examples.

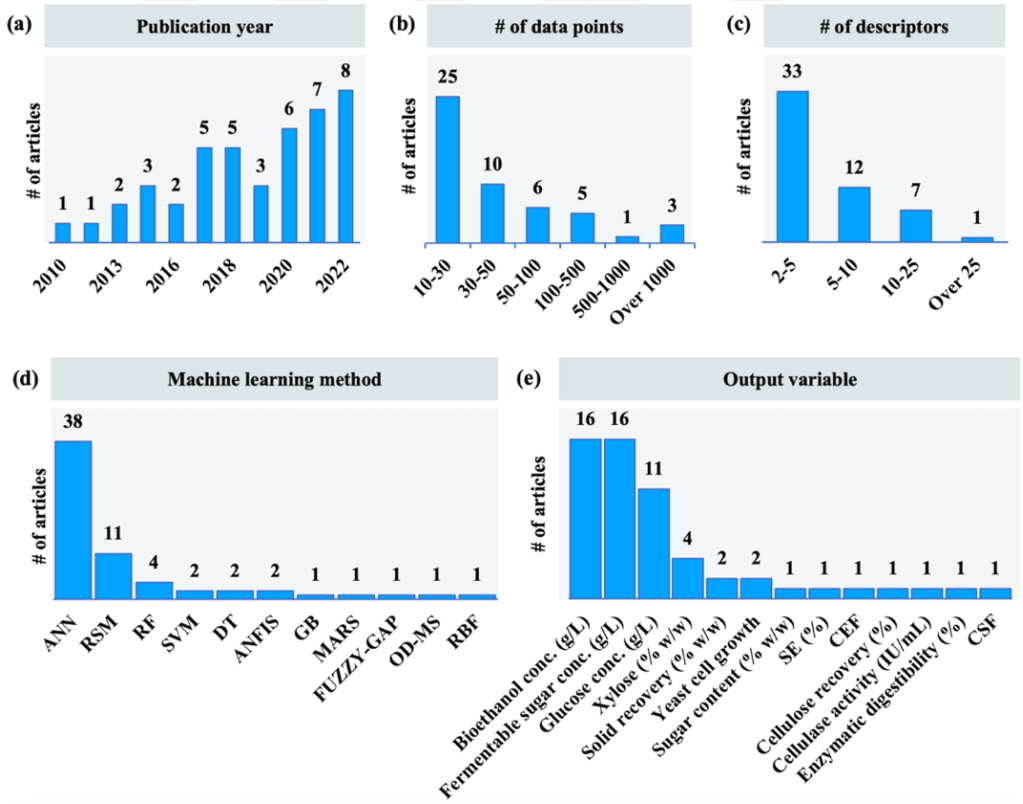

Figure 3.4 ML trends for lignocellulosic ethanol production in papers covered in this review: (a) distribution over publication dates; distribution of (b) data sizes; (c) number of descriptors; (d) ML algorithms used; and (e) output variables used in ML models [63].

In this section of this chapter, only the works directly related to lignocellulosic bioethanol production through the fermentation route are covered. For consistency, academic databases (i.e., Web of Science, Scopus, and Google Scholar) are 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.

Figure 3.4 summarizes the articles presented in this work with the numbers in the figure denoting the number of articles. Figure 3.4a shows the publication years of the articles, which shows an increasing trend in machine learning studies in the lignocellulosic bioethanol field even though there are oscillations in the trend due to the small size of data. The distribution of data size used in these works is given in Figure 3.4b. It is indicated that the majority of the studies have data sizes ranging from 10 to 30 data points, which is likely due to the time-consuming nature of experimental works in the field. As a consequence of the small data size, the number of descriptors is also small (*i.e.*, 2 to 5) in many works as given in Figure 3.4c; 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 3.4d shows the choice of ML algorithm in the studies. The output variables are also given in Figure 3.4e; although most of the 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 Table 3.4, Table 3.5 and Table 3.6 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 3.4. On the other hand, articles that include inputs from the hydrolysis step, but exclude the fermentation step are

summarized in Table 3.5. In Table 3.5, studies are categorized into; models that include variables from pretreatment and hydrolysis steps, and models only include hydrolysis step variables. In Table 3.6, articles that include the fermentation step inputs into the ML models are summarized with categorization performed in the same way as in Table 3.5.

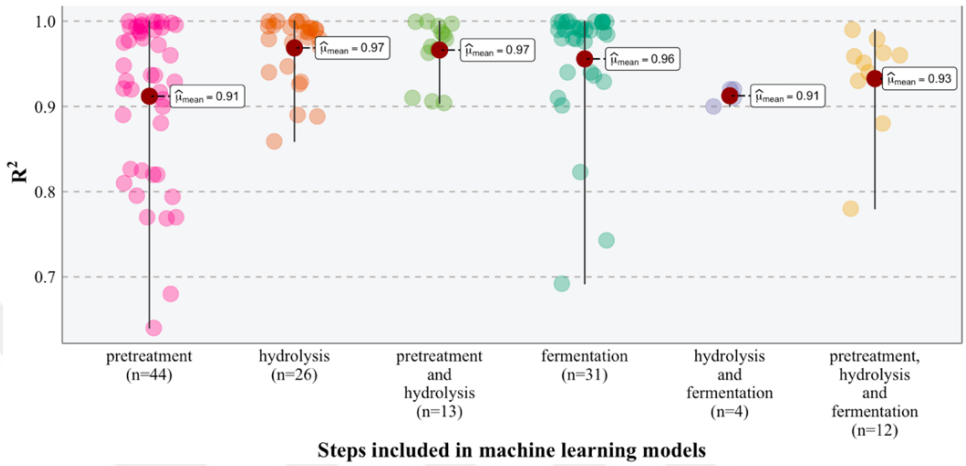

Figure 3.5 R^2 values of ML models with respect to corresponding steps of the input variables used in the models (n: number of models, $\hat{\mu}_{mean}$:average of the R^2 values) [63].

For a comprehensive picture, the unitless performance metrics (i.e., R^2) of all ML models studied in the research articles mentioned, are summarized in Figure 3.5. The average performance of models is above R^2 value of 0.90, which suggests 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 3.6. The results are separated with respect to lignocellulosic biomass used in the models, as the bioethanol, fermentable sugar, and glucose concentrations are highly dependent on the nature of the feedstock.

As the initial phase in the production of ethanol from lignocellulosic biomass, 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.* 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 [264]. It was shown that RF has higher predictive performance for CEF and SR (% w/w), while the other ML algorithms performed better for CEF [264]. Similarly, Parkhey *et al.* investigated the effect of pretreatment conditions on cellulose recovery using rice straw by ANN models with the Levenberg-Marquardt back-propagation algorithm [265].

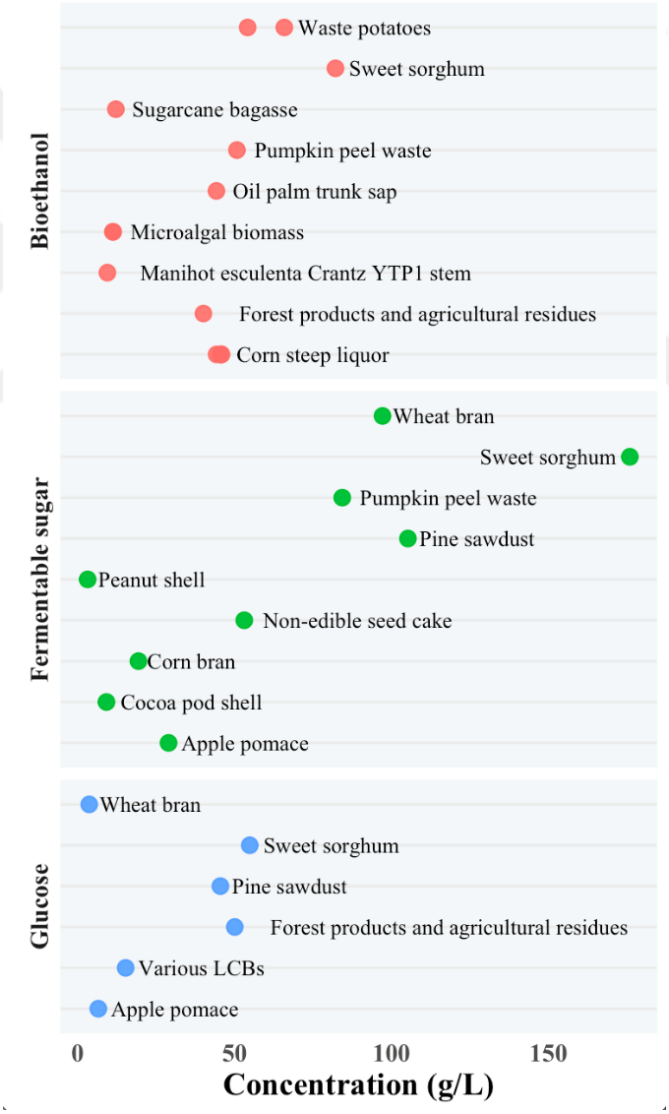

Figure 3.6. Prediction results of maximum concentration of selected outputs through ML-assisted modeling and optimization [63].

The cellulose content of lignocellulosic biomass is transformed into fermentable sugars by hydrolysis. Among the studies focused on the conversion efficiency of lignocellulosic biomass to fermentable sugars by machine learning models, some of them studied pretreatment and hydrolysis steps together. For example, Aruwajoye *et al.* [266] 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 [266]. Lee *et al.* and Moodley *et al.* 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 ^{o}C resulting in 356 mg/g biomass [267, 268], while Onay [269] offered ANN and RSM models for carbohydrate concentration; Chen *et al.* [270], Gitifar *et al.* [271], and Plazas Tovar *et al.* [272], on the other hand, modeled the glucose concentration as the output variable. Likewise, Ethaib *et al.* [273] studied both glucose and xylose as fermentable sugars, whereas Chang *et al.* [274] modeled enzymatic digestibility (%) by using inputs from both pretreatment and enzymatic hydrolysis steps. Charte *et al.* [70] 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). For example, Rivera *et al.* [275] studied glucose yield, while Vani *et al.* [84] studied both glucose and xylose yield; both used ANN as the ML algorithms. Gama *et al.* [276], and Giordano *et al.* [277], also modeled glucose and fermentable sugar concentration together. On the other hand, various researchers modeled fermentable sugar concentration and determined its maximum as 3.10 g/L [278], 84.27 g/L [279], 9.10 g/L [280], 53.03 g/L [281], 0.704 g/g [282] and 175.94 g/L [283]. As an additional example, Parkhey *et al.* [265] studied the effects of enzymatic hydrolysis process variables on saccharification efficiency (SE) by using ANN.

Table 3.4. Summary of machine learning on lignocellulosic bioethanol production including 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	(23) biomass characteristic, pretreatment condition, main ILS identity, co-ILS identity, catalyst loading	SVM	R^2= 0.8998 , RMSE= 0.2808	[264]
					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	Cellulose recovery (%)	20	(3)	ANN	R^2= 0.999[a], 0.999[b], 1[c], 0.993	[265]
Water hyacinth	Dilute-acid pretreatment	Xylose (mg/g)	30	(4)	ANN	R^2= 0.9995	[287]
Mixed vegetable waste	Dilute-acid pretreatment	Fermentable sugar conc. (mg/g)	29	(3) reaction time, reaction temperature, acid conc.	ANN	R^2= 0.92-0.98	[288]
			58		RSM	R^2= 0.7686-0.9368	
Pennisetum grass	Alkali pretreatment; NaOH	Fermentable sugar (g/g)	16	(3) pretreatment temperature, acid conc., soaking time	ANN	R^2= 0.98 , RMSE= 0.027	[289]
		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	[290]
					ANN	R^2= 0.9933 , RMSE= 1.129	
Hardwood (L. leucocephala)	Organosolv treatment; glycerol	Fermentable reducible sugar (g/g)	17	(3) catalyst conc., duration, temperature	RSM	R^2= 0.996 , RMSE= 5.564	[291]
			17*3		ANN	R^2= 0.998 , RMSE= 3.630	
Wheat straw	Dilute-acid pretreatment	Xylose yield (%)	17	(3) temp., time, acid conc.	GPR	R^2= 0.999	[82]
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	[266]
					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]	[70]
					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	[270]
					MARS	R^2= 0.929	
Napiergrass	Steam explosion followed by alkali pretreatment; NaOH	Enzymatic digestibility (%)	27	(3) steam explosion (temp., time, and particle size)	ANN	R^2= 0.988[a], 0.975[b]	[274]
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	[273]
		Xylose (% w/w)				R^2= 0.9479	
Oil palm empty fruit bunches	Ultrasonic assisted organosolv treatment;	Fermentable sugar (g/g)	20	(3) temperature, time, sonication power	ANN	R^2= 0.90843[a], 0.8264[c]	[267]

[a] performance measures for train set, [b] performance measures for test set, [c] performance measures for validation set, [d]

Table 3.5. Summary of machine learning on lignocellulosic bioethanol production including hydrolysis step.

Lignocellulosic biomass	Pretreatment	Hydrolysis	Output variable [d]	Data size	Descriptors (# of descriptors)	ML method [e]	Performance Measures [f]	Reference
pretreatment and hydrolysis								
Sugarcane leaf waste	Inorganic salt-based treatment; $ZnCl_2$	enzymatic hydrolysis	Fermentable sugar yield (% w/w)	90	(8)	ANN	R^2= 0.97	[268]
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= 0.999[a], 0.994[c], 0.999[b]	[269]
						RSM	R^2= 0.91. MSE= 0.78	
Sugarcane bagasse	Dilute-acid pretreatment; H_2SO_4	enzymatic hydrolysis	Glucose yield (%)	3049	(6)	ANN	MSE= 6.8, R^2= 0.987	[272]
Lignocellulosic biomass	Dilute-acid pretreatment	enzymatic hydrolysis	Fermentable sugar conc. (g/L)	107	(9) biomass comp., pretreatment cond., cellulose to lignin, cellulase conc., pretreatment	ANN	R^2= 0.997[a], 0.984[c], 0.967[b]	[292]
						PRM	R^2= 0.963	
Corn stover	Dilute-acid pretreatment; HCl, H_2SO_4, H_3PO_4	enzymatic hydrolysis	Phenolic contents and glucose yield	77	(6) acid conc., pretreatment temp., residence time, solid-to-liquid, acid conc., enzyme dosage	ANN	R^2= 0.904(phenolic conc.) 0.906 (glucose conc.)	[293]
hydrolysis								
Pumpkin peel waste	-	enzymatic hydrolysis	Fermentable sugar conc. (g/L)	30	(4)	ANN	R^2= 1[a], 0.99979[c], 0.99988[b]	[279]
						RSM	R^2= 0.988	
Cocoa pod shell	Microwave	acid hydrolysis	Fermentable sugar conc. (g/L)	12	(2) cocoa pod shell weight, H_2SO_4 conc.	RSM	R^2= 0.89	[281]
						ANN	R^2= 0.94	
Non-edible seed cake	Autoclave	acid hydrolysis	Fermentable sugar conc. (g/L)	12	(2) time, HCl conc.	ANN	R^2= 0.975, RMSE= 1.078	[280]
						RSM	R^2= 0.888, RMSE= 2.139	
Waste broken rice	-	enzymatic hydrolysis	Fermentable sugar yield (g/g)	30	(4) temperature, time, pH and enzyme dosages	ANN	R^2= 0.993, RMSE= 0.078	[282]
						RSM	R^2= 0.987, RMSE= 0.102	
Peanut shell	Combination of alkali, dilute acid, steam explosion	enzymatic hydrolysis	Fermentable sugar conc. (g/L)	45	(3) temperature, substrate conc., and spore dosage	ANN	R^2= 0.929	[278]
Rice straw	Microwave-assisted alkali treatment; NaOH	enzymatic hydrolysis	SE (%)	30	(4) substrate conc., enzyme load, temperature, and Tween-80 conc.	ANN	R^2= 0.991[a], 0.926[b], 0.981[c], 0.947	[265]
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]	[84]
Apple pomace	-	enzymatic hydrolysis	Glucose and fermentable sugar	81	(4) substrate loading, enzyme loading, temperature, initial pH	ANN	R^2= 0.99	[276]
Corn bran, wheat bran and pine sawdust	-	acid hydrolysis	Glucose conc.	70	(4) hydrolysis temp., H_2SO_4 conc., acid solution/feedstock ratio, hydrolysis time	RBF-PSO	R^2= 1.000, 1.000 and 0.995	[277]
			Fermentable sugars conc. (g/L)				R^2= 0.979, 0.859 and 0.992	
Sweet sorghum	-	enzymatic hydrolysis	Fermentable sugar conc. (g/L)	29	(4) substrate loading, α-amylase conc., amyloglucosidase conc., stroke speed	ANN	R^2= 0.994	[283]
Sugarcane bagasse	Alkali pretreatment; H_2O_2	enzymatic hydrolysis	Glucose yield (% theoretical max.)	480	(3) cellulase, β-glucosidase, time	ANN	acceptable performance	[275]

([a] performance measures for train set, [b] performance measures for test set, [c] performance measures for validation set, [d] SE: saccharification efficiency, [f] RMSE: Root mean square error)

Table 3.6. Summary of machine learning on lignocellulosic bioethanol production including the fermentation step.

Lignocellulosic biomass	Pretreatment	Hydrolysis	Fermentation	Output variable	Data size	Descriptors (# of descriptors)	ML method [d]	Performance Measures [e]	Reference
pretreatment, hydrolysis and 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]	[94]
							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	[284]
							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 and fermentation									
Forest products and agricultural residues	-	enzymatic hydrolysis	*S. Cerevisiae*	Glucose yield (g/L)	300	(11) biomass composition; saccharification time, temperature, pH, shaking speed; fermentation time, temperature, pH, shaking speed	OD-MS	overall accuracy is 95%	[285]
				Bioethanol yield (g/L)					
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]	[286]
							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)	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	[294]
Intermediates and byproduct of sugar beet processing	*not necessary, not lignocellulosic	-	*S. Cerevisiae*	Bioethanol content (% v/v)		(3) fermentation time, starting sugar content, substrate type	ANN	R^2 from 0.823 to 0.999	[295]
				Yeast cell number (10^8 cells ml/L)				R^2 from 0.692 to 0.993	
				Sugar content (% w/w)				R^2 from 0.929 to 0.999	

([a] performance measures for train set, [b] performance measures for test set, [c] performance measures for validation set, [e] RMSE: Root mean square error,)

Table 3.6. Summary of machine learning on lignocellulosic bioethanol production including the fermentation step (cont.).

Lignocellulosic biomass	Pretreatment	Hydrolysis	Fermentation	Output variable	Data size	Descriptors (# of descriptors)	ML method [d]	Performance Measures [e]	Reference
Fermentation									
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]	[279]
							RSM	R^2= 0.9762	
Corn cobs and corn stovers	-	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	[296]
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	[297]
Corn steep liquor	-		*S. Cerevisiae*	Bioethanol prod. (g/L)	46	(5) substrate conc., pH, time, temperature, inoculum size	SM	R^2= 0.34, RMSE= 1.29	[298]
							NN	R^2= 0.98, RMSE= 0.19	
			Instant dry yeast		46	(5) substrate conc., pH, time, temperature, inoculum size	SM	R^2= 0.98, RMSE= 0.97	
							NN	R^2= 0.99, RMSE= 0.29	
Watermelon waste	-	enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol prod. (% w/w)	27	(2) agitator speed, yeast amount	NN	R^2= 0.9895	[90]
							NFIS	R^2= 0.9993	
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	NN	R^2= 0.99971[a], 0.74285[c], 0.93635[b], 0.90114	[269]
							SM	R^2= 0.94	
Manihot esculenta Crantz YTP1 stem	Dilute-acid pretreatment; CH_3COOH, HNO_3	enzymatic hydrolysis	*Z. mobilis*	Cellulase activity (IU/mL)	30	(4) pH, temperature, agitation and time	NN	R^2= all 0.990, MSE= 0.2654, RMSE= 0.5151	[299]
				Bioethanol yield (g/L)				R^2= all 0.979, MSE= 0.4324, RMSE= 0.6575	
Oil palm trunk sap	-		*S. Cerevisiae*	Bioethanol conc. (g/L)		(4) fermentation time, pH, temperature, total sugar	NFIS	R^2= 1[a], 0.9991[b], 0.99975[c]	[300]
Breadfruit starch hydrolysate	-		Instant dry yeast	Bioethanol yield (% v/v)	17	(3) reducing sugar conc., fermentation time, pH	NN	R^2= 0.9995	[301]
Sweet sorghum	-	enzymatic hydrolysis	*S. Cerevisiae*	Bioethanol conc. (g/L)	17	(3) yeast conc., reaction temperature, agitation speed	NN	R^2= 0.987	[283]

Fermentation, as the final phase of the three-step process of bioethanol production from lignocellulosic biomass, has been modeled by itself or together with the previous steps by various researchers. For example, Smuga-Kogut *et al.* [94], and Suresh *et al.* [284] modeled bioethanol production by using the variables belonging to three consecutive steps as the descriptors. Suresh *et al.* [284] 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 is superior in modeling bioethanol production. Smuga-Kogut *et al.* [94] used three different ML methods as 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 Vinitha *et al.* [285], and Fischer *et al.* [286].

A large number of groups modeled the fermentation step only with the output variable of bioethanol concentration (all used different forms of ANN). On the other hand, Konishi studied cell growth and bioethanol production together [296], while Selvakumar *et al.* investigated cellulase activity and bioethanol yield [299]. Grahovac *et al.*, on the other hand, analyzed the bioethanol content, yeast cell number, and sugar content in the fermentation process [295].

Construction of a sufficiently large and accurate dataset is one of the biggest challenges for machine learning applications in many fields; this also seems to be the case for lignocellulosic ethanol production. In fact, this may be more problematic for complex systems, such as lignocellulosic ethanol production, because a larger number of descriptors is required to adequately represent such systems, 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 became the dominant route. Considering that the descriptors are different for dissimilar routes, data from different process combinations are different. 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 that are tested the first time create unique variables that are not reported by other papers. All these create significant difficulties for the implementation of machine learning, which relies on learning from existing relations in the data set; single or few data points having variables that are not shared by the others have limited use in machine learning analysis. The data seems to be a bigger problem for more complex configurations like performing hydrolysis and fermentation at the same time 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 in especially pretreatment and hydrolysis steps [302]. Normally this should not be a problem for machine learning 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 associated with the descriptors.

3.3. Biochar Production

3.3.1. Bibliometric and Keyword Analysis of Biochar Production Studies

Bibliometric analysis was conducted on publications indexed in the Web of Science database using the keyword "biochar" as the focal point, resulting in a total of 34,876 publications. (by Feb 20, 2024). Author keywords are extracted from the publications and examined to discern trends in the field. Analysis began in 2012 due to the insufficient number of biochar-related keywords in earlier years. Also, three-year moving averages are applied to mitigate yearly fluctuations. (Figure 3.7).

As observed, pyrolysis is the dominant production technology (Figure 3.7a); although traditional pyrolysis methods (*e.g.*, slow pyrolysis) are more common in the early years, advanced techniques like *co-pyrolysis*, *microwave-assisted pyrolysis* (MAP) and *catalytic pyrolysis* seem to be getting more interest in the recent years. Feedstock used for biochar

production is mainly categorized as crop and forestry residues, with straw and sugarcane bagasse as the most common feedstock while wastewater sludge, municipal waste, manure, and algae are also used as the keywords (Figure 3.7b). Response surface methodology, which represents the traditional experimental design and modeling approach, is still the most common method for modeling while data-driven methods (*e.g.*, ANN) have increased in recent years (Figure 3.7c).

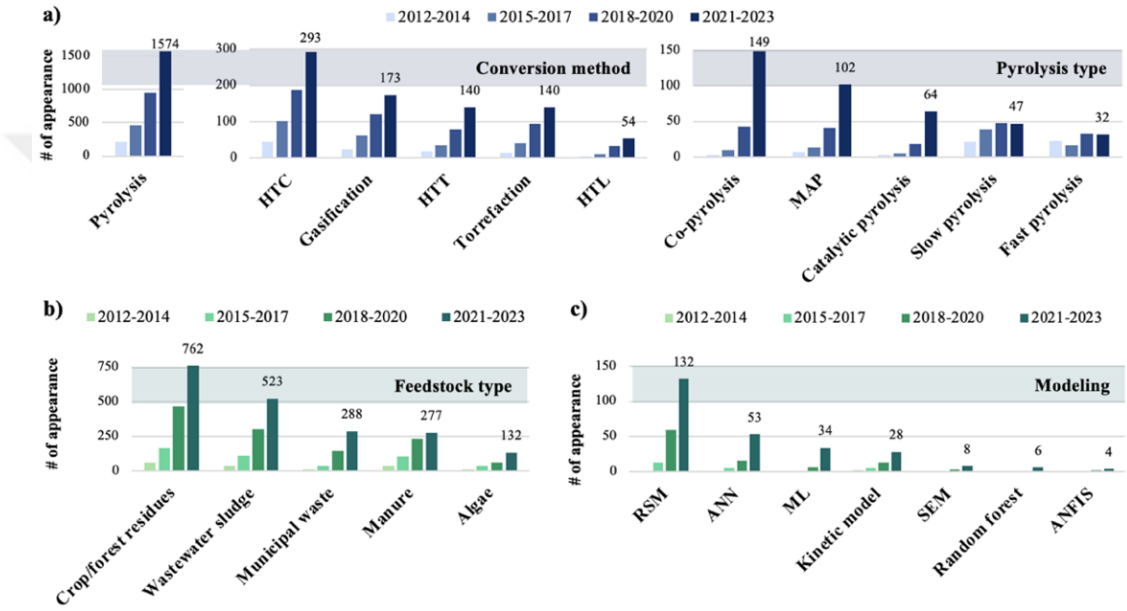

Figure 3.7. Publications trends in biochar research through the years: (a) production processes; (b) biomass source; (c) modeling.

3.3.2. Machine Learning on Biochar Production

Due to the complexities in the production processes, it is hard to determine the best possible feedstock and process conditions to produce biochar with the desired properties and the optimal economic cost [303]. For example, pyrolysis, is a complex process involving a wide range of reactions, including dehydration, depolymerization, isomerization, aromatization, decarboxylation, and charring [124]. In addition, pyrolysis highly depends on the feedstock composition (*e.g.*, moisture, volatile matter, and ash content) and operating variables (*e.g.*, temperature, pressure, residence time, and heating rate). Consequently, variations in process variables usually cause notable variations in final product composition.

Machine learning (ML) may come into play here and can help determine the combination of variables that leads to the highest biochar yield with desired properties [304]. It can be used to extract knowledge from data by using algorithms such as ANNs, SVMs, RF regression, and DTs to perform specific tasks such as prediction, classification, clustering, and association [186]. In this section of this chapter a detailed review and analysis of ML applications in biochar production is summarized.

The objective and coverage of ML applications in biochar production are quite diverse. It is found that most of the articles compare the performance of different ML algorithms for the prediction of biochar yield or some other properties [305-308]. Other research areas include the study of ML model performance by change of input variables using the same ML models [100, 247, 309, 310], the predictive power of ML models for different conversion methods [311], and the effect of different optimization methods on ML performance [312-314].

The desired outcome of the biochar production process can be summarized in two groups, which are also used in ML applications as target variables. First, biochar production should be carried out efficiently; hence, variables like biochar yield, or raw material utilization have to be optimized as the target variable. Second, the biochar produced should have certain properties depending on the application type; hence, properties like surface area, porosity, surface functional groups, and elemental composition may need to be also optimized as the output.

Inputs used in the ML models to predict the biochar yield and biochar properties can be grouped into four categories; reaction conditions (temperature, heating rate, time, *etc.*), biomass content (contents of ash, volatile matter, fixed carbon, *etc.*), the elemental composition of biomass (carbon, hydrogen, oxygen, nitrogen, *etc.*), and structure of the biomass (cellulose, hemicellulose, and lignin contents). Biomass used in ML studies covers a large number of raw materials, including energy crops, crops/forest residues, animal manure, food waste, algae, sewage sludge, biodegradable organic wastes, and their mixtures. The majority of the ML models reviewed in this article used pyrolysis for biomass production as represented in Table 3. 7; however, the hydrothermal treatment methods such

as hydrothermal liquefaction, hydrothermal carbonization, and torrefaction are also represented in Table 3. 8.

The target (output) variables to predict or optimize are quite diverse as presented in Figure 3.8a. While the *biochar yield* is the most commonly used output, as the key performance indicator for the production process, the output variables that are relevant for specific applications are also used. For instance, biochar can be used as fuel (energy storage) or utilized for carbon capture and sequestration, soil amendment, catalyst, material for electrodes, and pollutant removal. Since each of these applications requires different biochar properties, which should be optimized during biochar production, the output variables for ML applications are also different. Indeed, in addition to the yield, the output variables like carbon content (C-char), higher heating value (HHV), energy recovery efficiency (ER), and energy densification (ED) are also modeled if biochar is used as an alternative energy source [306]. Carbon capture and storage stability of biochar, on the other hand, is commonly quantified by biochar properties of C-char and carbon recovery (CR). SSA of biochar is an important property that determines its performance in some specific applications [315]. Together with other properties like total pore volume (TPV) as well as surface functional groups (*e.g.*, N-containing groups), biochar surface characteristics are decisive for its use as an adsorbent for reducing greenhouse gas emission, removal of metals or organic pollutants, as well as electrode materials for desalination through capacitive deionization and mitigate climate change [305, 316].

ANN, RF, and SVMs are the most commonly employed ML algorithms (Figure 3.8b). ML models are also diversified by some research groups by including other less frequently used algorithms such as adaptive neuro-fuzzy inference system (ANFIS), multivariate adaptive regression splines (MARS), Gaussian process regression (GPR), and ensembled learning tree (ELT). Various optimization methods are also employed complementary to ML models. Although particle swarm optimization (PSO) is the most common one; artificial bee colony technique (ABC), flower pollination algorithm (Flower), genetic algorithm (GA), grasshopper optimization algorithm (GOA), grey wolf optimization (GWO), whale optimization algorithm (WOA), Rao algorithms (RA), simulated annealing optimization

(SA), and sine cosine algorithm (SCA) are other methods that are used for hyperparameter tuning.

The major descriptors used in the reviewed articles are presented in Figure 3.8c. It is indicated for biochar yield models that; reaction conditions are the most commonly used set of descriptors. However, taking a property or process variable as a descriptor in an ML model does not necessarily mean that it will be really deterministic in model predictions; feature importance analysis should be performed for this as done in most of the papers reviewed here. The relative importance of features can be different for different data sets and algorithms used; hence, it is not possible to construct a single list that is valid for all works reviewed. Nevertheless, some generalizations still can be made; for example, we can see that the *reaction temperature* is the most frequently repeated descriptor followed by the ash and carbon content of the feedstock. For HHV prediction, on the other hand, proximate analysis of the feedstock is the most commonly employed descriptor in the models (ash content is the most common) as also found in feature importance analysis In C-char models, retention time becomes important as the time required for the reaction has a very high impact on carbonization while N-char prediction is highly dependent on the nitrogen content of the feedstock as expected.

Figure 3.8d summarizes the articles that examine the effect of ML methods on the prediction of biochar yield. Among the studied models, those developed using LR and MLR are the least accurate, and all ML models demonstrate superior predictive performance compared to those traditional methods. Notably, models employing GB and XGB exhibit the highest accuracy, followed closely by those using RF algorithms.

The studies reviewed in this work can be divided into two groups with respect to the dataset used in the models. As some used the experimental data they created in their laboratory, others curated datasets from results published in the literature. As expected, the first category comprises a smaller number of data points apparently due to the difficulties in performing experiments. The latter group has the advantage of creating larger datasets that are more suitable for complex processes with a large number of input variables; however, the variations of feedstock and lack of standard protocols for experimenting and reporting

the results make these datasets noisier than those created in a single laboratory. It is fair to say that most of the reviewed articles reported high R^2 values (usually higher than 0.8) for the plot of ML predicted values of target variables against experimental values, and this level of fitness can be considered as satisfactory for predictive models. There seems to be no distinct advantage of any ML algorithms probably due to the continuous nature of descriptors; as far as we know, most of the algorithms work well with continuous data as in the biochar production case. Furthermore, it is not easy to compare the fitness indicators (*e.g.*, R^2) and generalize because the distinction among the fitness of training, validation, and testing sets is not always made during reporting the results. Also, no clear change in model fitness is observed with the change in data size either.

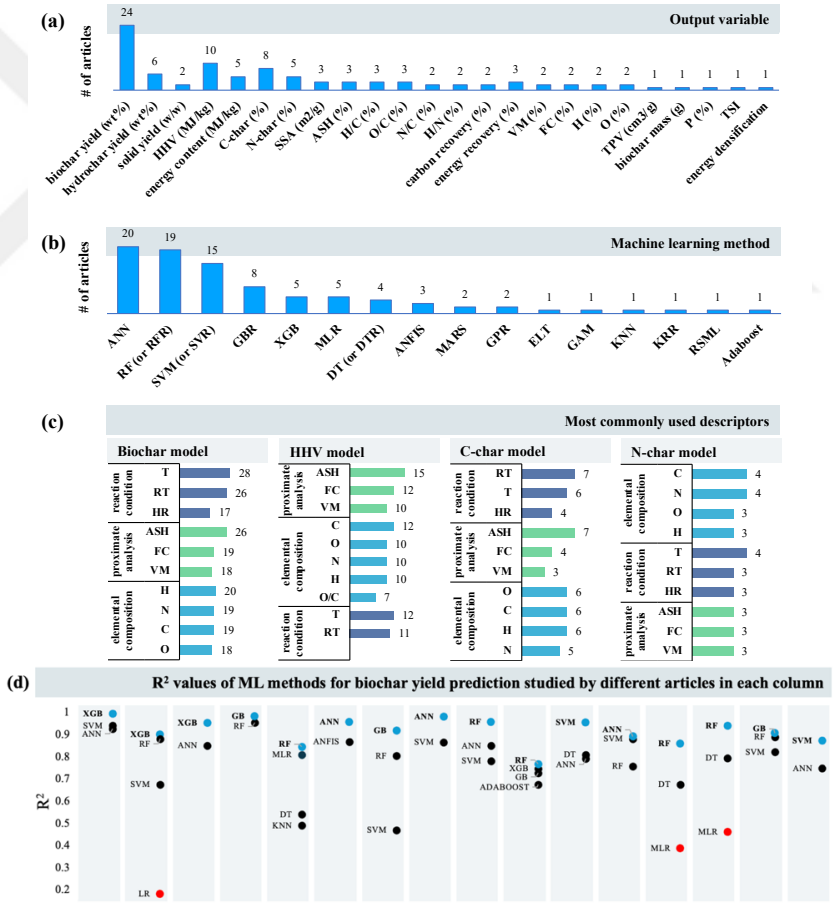

Figure 3.8. Machine learning application trends in biochar production articles covered in this review. Distribution of articles (a) output variables used (b) machine learning algorithms used, (c) frequency of descriptors used in different models, (d) R^2 values of ML models compared by different articles given in each column for biochar yield.

Table 3.7. Summary of machine learning papers on biochar production via pyrolysis.

Data Size	(# of Descriptors) Descriptors	Target Variable	ML Algorithms	Reference
482	(5) structural analysis (LIG), proximate analysis (ASH, PS), reaction conditions (T, HR)	biochar (wt%)	ANN	[310]
245	(12) structural analysis (CEL, HC, LIG), ASH, PS, elemental composition (C, H, O, N), reaction conditions (HR, HTT, RT)	biochar (wt%)	RF	[100]
128		C-char (%)		
-	(4) proximate analysis (ASH, FC, VM), T	biochar (wt%)	XGB	[247]
122	(8) structural analysis (CEL, HC, LIG), ASH, PS, reaction conditions (T, GFR, FR)	biochar (wt%)	RF (XGB, SVM, MLP)	[317]
538	(7) elemental composition (C, H, O, N, molar ratio of O/C, H/C, N/C), ASH, reaction conditions (T, HR, RT)	biochar (wt%)	RF	[318]
276		energy content (MJ/kg)		
305		C-char (%)		
276		N-char (%)		
419	(6) reaction conditions (T, RT, capacity), proximate analysis (PS, FC, ASH)	biochar (wt%)	ANN	[319]
422		FC char (%)		
422		VM char (%)		
422		ASH char (%)		
393		HHV (MJ/kg)		
18	(8) structural analysis (CEL, HC, LIG), proximate analysis (ASH, FC, VM, MC), T	solid, liquid and gas product yields	ANN	[320]
72	(11) reaction conditions (T, HR), elemental composition (C, H, O, N), proximate analysis (ASH, FC, VM, MC), HHV	biochar (wt%)	ANN	[309]
33	(5) reaction conditions (T, HR, RT, FR), MC	biochar (g)	LS-SVM (ANN)	[321]
-	(13) reaction conditions (T, RT, GFR, HR), proximate analysis (VM, ASH, FC), elemental composition (C, H, K, Ca, Mg, Fe)	biochar (wt%)	ANN	[322]
157	(11) proximate analysis (VM, ASH, FC), elemental composition (C, H, O, N), reaction conditions (capacity, MW power, RT, MW absorber)	biochar (wt%)	XGB (LR, RF, SVM)	[307]
67		HHV (MJ/kg)		
237	(10) proximate analysis (VM, ASH, FC), elemental composition (C, H, O, N), reaction conditions (T, HR, RT)	biochar (wt%)	RSML	[323]
197		C-char (%)		
113		HHV (MJ/kg)		
94	(16) composition of biomass (C, H, O, N) and plastics (C, H, O, N, Cl), proximate analysis (VM, ASH, FC), reaction conditions (T, RT, HR), plastic percent in the sample	biochar (wt%)	XGB (DNN)	[114]
152	(11) proximate composition (VM, ASH, FC), elemental composition (C, H, N, S, O), reaction conditions (T, RT, HR)	biochar (wt%)	GBR (RF)	[316]
169		SSA (m^2/g)		
169		TPV (cm^3/g)		
152		yield-SSA-TPV		
292	(9) proximate analysis (MC, VM), biomass composition (C, N), structural analysis (CEL, HC), reaction cond. (T, RT, HR)	biochar (wt%)	RF (MLR, DT, KNN)	[305]
	(9) proximate analysis (MC, ASH), biomass composition (C, H, N, S), PS, T, RT)	SSA (m^2/g)		

Table 3.7. Summary of machine learning papers on biochar production via pyrolysis (cont.).

Data Size	(# of Descriptors) Descriptors	Target Variable	ML Algorithms	Reference
226	(10) proximate analysis (VM, ASH, FC), elemental composition (C, H, O, N), reaction conditions (T, HR, RT)	biochar (wt%)	MLPNN (ANFIS)	[324]
159		FC char (%)		
159		VM char (%)		
159		ASH char (%)		
162		C-char (%)		
150		H content (%)		
150		O content (%)		
162		N-char (%)		
249	(14) elemental composition (C, H, O, N, S), proximate analysis (VM, FC, ASH), reaction conditions (T, MW power, RT, MW absorber percentage, dielectric constant of absorber, dielectric loss factor of absorber)	biochar (wt%)	GBR (SVR, RFR)	[325]
		H/C ratio in char (%)		
		H/N ratio in char (%)		
		O/C ratio in char (%)		
		energy content (MJ/kg)		
149	(10) elemental composition (C, H, O, N, molar ratio of H/C, O/C), proximate analysis (ASH, VM, FC, VM/FC)	HHV (MJ/kg)	GBR (RF, SVM)	[326]
165	(10) proximate analysis (FC, ASH, VM), elemental composition (C, H, O, N), reaction conditions (T, HR, RT)	-	SVR (RF)	[311]
244	(11) elemental composition (C, H, O, N, S), proximate analysis (VM, FC, ASH), reaction conditions (T, HR, RT)	biochar (wt%)	RFR (MLPNN, SVR)	[115]
		O/C ratio in char (%)		
		H/N ratio in char (%)		
		H/C ratio in char (%)		
		energy content (MJ/kg)		
645	(12) proximate analysis (MC, FC, VM, ASH), elemental composition (C, H, O, N), reaction conditions (T, HR, sweep flow rate, PS)	biochar (wt%)	RF (Adaboost, GBDT, XGB)	[327]
1626	(3) type of feedstock, T, RT	TSI	MARS (ANN)	[328]
443	(16) elemental composition (C, H, O, N), proximate analysis (ASH, FC, VM, MC), structural analysis (CEL, HC, LIG), reaction conditions (HR, T, PS, RT, N_2 flow)	biochar (wt%)	ELT-PSO (GPR, SVM, DT with GA, PSO)	[313]
33	(5) reaction conditions (T, RT, HR), MC, FR	biochar (wt%)	ANFIS-GWO (LS-SVM, ANN)	[312]
402	(9) proximate composition (FC, VM, MC, ASH), structural analysis (CEL, LIG), reaction conditions (HTT, RT, HR)	biochar (wt%)	ANN-RA (SCA, GA, PSO)	[314]
118	(13) elemental composition (C, H, N, O), proximate analysis (ASH, VM, FC), structural analysis (CEL, HC, LIG), reaction conditions (T, RT, HR)	biochar (wt%)	GBR (RF)	[315]
161	(10) elemental composition (C, H, N, O), proximate analysis (ASH, VM, FC), reaction conditions (T, RT, HR)	N-char (%)		
183		SSA (m^2/g)		
120		yield-char, N-char, and SSA		

Table 3.8. Summary of machine learning papers on biochar production via hydrothermal liquefaction, hydrothermal carbonization, and torrefaction.

Data Size	(# of Descriptors) Descriptors	Target Variable	ML Algorithms	Reference
81	(6) structural analysis (CEL, HC, LIG), MC, reaction conditions (T, RT)	hydrochar (wt%)	XGB (SVM, MLPANN)	[329]
296	(11) elemental composition (C, H, O, N), proximate analysis (ASH, FC, VM), reaction conditions (T, RT, solid–liquid ratio)	hydrochar (wt%)	ANN-PSO	[330]
		HHV (MJ/kg)		
		ASH char (%)		
		N/C ratio in char (%)		
248	(10) proximate analysis (FC, ASH, VM), elemental composition (C, H, O, N), reaction conditions (T, RT, MC)	-	SVR (RF)	[311]
108	(3) reaction conditions (T, RT, GFR)	biochar (wt%)	DLNN (SVM)	[201]
246	(11) proximate analysis (ASH, TS, OM), elemental composition (C, molar ratio of H/C, O/ C, N/C), HHV, reaction conditions (MC, T, RT)	HHV (MJ/kg)	RF (GBT, ANN)	[308]
		CR (%)		
		ER (%)		
248	(10) elemental composition (C, H, N, O), proximate analysis (FC, VM, ASH), reaction conditions (RT, T, MC)	hydrochar (wt%)	DNN (RF, SVR)	[306]
		HHV (MJ/kg)		
		C-char (%)		
		ER (%)		
		CR (%)		
		H/C ratio in char (%)		
		O/C ratio in char (%)		
		N/C ratio in char (%)		
around 800	(6) elemental composition (molar ratio of O/C, H/C), ASH, reaction conditions (T, RT, capacity)	hydrochar (wt%)	RF (MLR, DT)	[331]
		energy content (MJ/kg)		
		C-char (%)		
138	(5) elemental composition (C, N), proximate analysis (VM, FC), reaction conditions (T)	N-char (%)	ANN (MLR)	[332]
109	(12) proximate analysis (VM, FC, ASH), elemental composition (C, H, O, N, P), reaction conditions (T, RT, initial pH of the feedwater, dry matter content)	P content (%)	RF	[333]
475	(11) proximate analysis (ASH, FC), elemental composition (H, O), structural analysis (CEL, HC, LIG), polarity, solid conc., T, RT	hydrochar (wt%)	RF (MLR, DT)	[334]
	(12) proximate analysis (ASH, FC), elemental composition (C, H, O), structural analysis (CEL, HC, LIG), polarity, solid conc., T, RT	C-char (%)		
		energy content (MJ/kg)		
about 300	(6) proximate analysis (FC, VM), molar ratio of O/C, H/C, reaction conditions (T, RT)	HHV (MJ/kg)	SVM-SA (MARS, RF)	[335]
over 800	(10) proximate analysis (VM, ASH), elemental composition (C, H, N), PS, reaction conditions (RT, T, CO_2, and O_2 fraction in the reacting gas)	solid yield (w/w)	GBT (KRR)	[336]
410	(18) proximate analysis (VM, FC, ASH), elemental composition (N, S, molar ratio of H/C, O/C), structural analysis (CEL, HC, LIG), reaction condition (T, RT, N2, CO_2, O_2), reactor property (PS, sample mass loading, GFR)	solid yield (w/w)	GBT (RF, SVM)	[337]
		HHV (MJ/kg)		
652	(10) elemental composition (C, H, N, S, O), ASH, reaction conditions (T, pressure, RT, solid/water percentage)	hydrochar (wt%)	GPR (ANN, GAM, SVR)	[338]

Biochar production has been one of the most commonly investigated subfields of biotechnological research in recent years; in addition to its traditional roles as a critical fuel for some parts of the world, biochar is a critical agent for environmental remediation. The number of ML applications in biochar production, as well as its utilization has been also increasing, and this trend seems to increase in the future if the two main challenges in biochar production are considered. One of the biggest challenges in biochar production is the nonstandard nature of raw materials (and processes) used. The type of raw material (*e.g.*, wood) and its content may vary depending on the location, time, or even batch of runs. In such a process, it will be hard to determine the optimum conditions to be used for a long time. Instead, one can use large datasets containing the variables and performance relations under various conditions and determine the best set of optimum conditions for the new case using ML algorithms. The second important issue in biochar production is that the required biochar properties may change depending on the field of use. Such that, apart from the obvious differences between the biochar properties to be used as fuel, soil amendment agent, or material for environmental remediation, even the different impurities or pollution required in environmental cleaning may require different biochar properties. Indeed, *functionalized* or *engineered* biochar [339], which has been studied extensively in recent years, creates a big opportunity for environmental remediation in the future considering the diversity of pollutants in type and concentrations in different wastewater sources. ML can provide a significant contribution to the solution of these challenges in two ways. First, it may be used for the *characterization* tests to establish the relation between the process conditions and biochar properties like surface area, porosity, and surface morphology as well as the quality and degree of functionalization (or modification in general); ML contribution will be more apparent with the increasing resolution (and therefore data size and information content) of the characterization techniques. Second, ML will contribute to the improvement of functionalized or engineered biochar by optimizing the production and utilization steps simultaneously. The *functionalization* is performed during the *production* and *pretreatment* of biochar but tested in *utilization*; hence, the data for all these steps has to be merged and analyzed together so that the descriptors (variables) of production and pretreatment steps could be optimized to have the desired performance in utilization.

4. ML ANALYSIS OF LIPID PRODUCTION FROM OLEAGINOUS YEASTS

This part of dissertation aims to investigate the relation of lipid content and cell growth in oleaginous yeasts and the factors that promote high lipid production using *Yarrowia lipolytica* by ML algorithms. For this purpose, a general dataset is constructed with major yeast strains, and a preliminary investigation is carried, and another dataset is constructed with extracting data from the published papers on biomass and lipid production from *Yarrowia lipolytica*, which was analyzed using decision tree classification to identify heuristic rules (as the combination of factors leading to high biomass production and lipid content) and association rule mining to determine the effects of individual factors on the same outputs.

4.1. Preliminary Analysis on Oleaginous Yeast Data

For preliminary analysis purposes, a more comprehensive dataset is constructed, with experimental data from different oleaginous yeast species. Most of the data comes from *Yarrowia lipolytica*, *Rhodosporidium toruloides*, *Lipomyces starkeyi*, *Cryptococcus curvatus*. In the articles that the data is gathered from, variables studied for their effects on oleaginous yeast growth and lipid accumulation is also studied. The dataset composed from 132 articles, 182 experiments, and 1367 data points. The most commonly studied variable was the effect of different carbon sources, followed by the carbon concentration and C/N ratio. The full list is given in Figure 4.1.

The interactions among three output variables were analyzed. According to the literature, yeasts exhibit higher growth rates but limited lipid content under optimal growth conditions and unlimited feed sources. Conversely, yeasts grown under stress conditions achieve higher lipid content but have lower growth rates. These behaviors are contradictory when aiming for high lipid production. The potential of an oleaginous yeast to achieve high lipid productivity is determined by both its growth rate and lipid content. To evaluate the

behavior of various species, a polynomial regression was fitted to the gathered data points. As shown in the Figure 4.2, the polynomial regression fit demonstrates high accuracy and reveals interesting trends, where all the data points are close to the regression surface. Specifically, when lipid content is lower, yeasts exhibit higher growth rates, whereas higher lipid content is associated with limited cell growth to a certain value. This dataset includes both genetically engineered yeasts and those grown under stress conditions, yet it reflects a general trend in oleaginous yeast growth. This trend can serve as a useful criterion for selecting potential yeast species that fall above the regression line or in the region indicating higher lipid production.

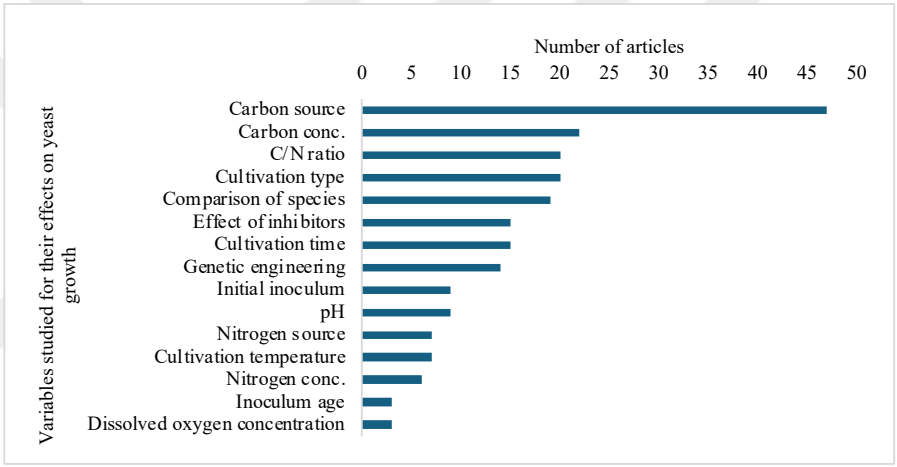

Figure 4.1. Number of publications studied the effect of different variables.

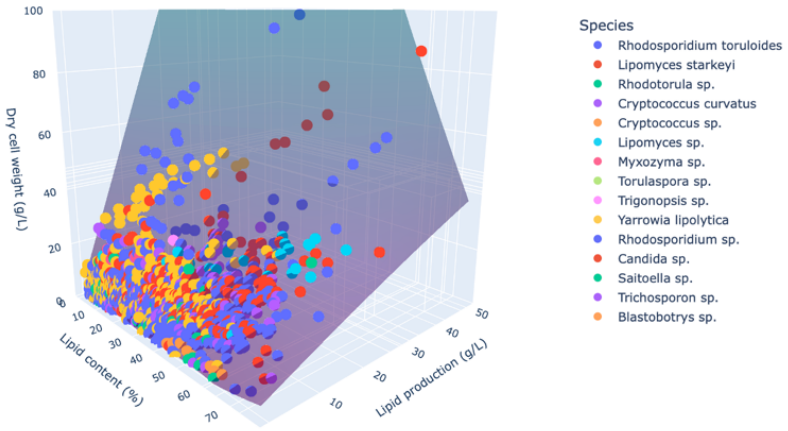

Figure 4.2. Relation of lipid content, lipid production, and dry cell weight of oleaginous yeast data in the dataset.

4.2. Machine Learning on *Yarrowia lipolytica* Dataset

The dataset described in Section 4 is used for ML algorithms. The specific computational details are presented in Section 4.2.1. Two ML algorithms with different tasks are employed; DT analysis to unveil specific rules to achieve high lipid production with respect to defined descriptors (Section 4.2.2), and ARM algorithm to analyze the effect of single descriptor effect on selected output variables (Section 4.2.3).

4.2.1. Computational Details

Yarrowia lipolytica database contains 356 instances (after eliminating entries containing unique variables or variables with missing information) from the experimental results reported in 22 articles (published from the year 2009 to 2020) [138, 144, 147, 340-358]. On the basis of relevance searches, the publications were extracted from major online sources (such as ACS, RSC, Elsevier, and Wiley). The dataset includes ten different descriptors (input variables) as well as three output variables, which are biomass production (g/L), lipid content (%) and lipid production (g/L). Table 4.1 lists the ten different descriptor variables used in the analyses, as well as their ranges or identities.

Table 4.1. Descriptors used and their details.

Input variable	**Ranges for continuous variables or identities for discrete variables**
strain type	A-101, ACA-DC50109, ACA-YC 5033, DiSVA 347, DSMZ No.3286, E26E1, H-222, LFMB19, LFMB20, M7, SKY7, SM7, W29
fermenter type	flask, batch reactor, fed-batch reactor
carbon sources and maximum concentrations ($g.L^{-1}$)	crude glycerol (168), pure glycerol (103), glucose (160), fructose (100)
initial C/N molar ratio	0-283
nitrogen sources and maximum concentrations ($g.L^{-1}$)	peptone (9.5), yeast extract (13.4), $(NH_4)_2SO_4$ (8.0), NH_4Cl (0.5)
variables related to oxygenation	sufficient oxygen, high oxygen, air support
temperature (°C)	15-45
pH	2-8
mixing speed (rpm)	100-800
time (h)	0-375

The apriori algorithm from the R package “arules” was used to do association rule mining [359]. The lift values of the association rules were used for knowledge extraction.

These values are interpreted by looking at their distance from "1", where a negative distance indicates a negative relationship and a positive distance indicates a positive relationship, with high values having more influence. The meaning of lift and the other two measures (support and confidence) are explained with examples in Section 4.2.3. Two alternative strategies were used to build the data during the discretization of the outputs. First, the output variables were discretized into two classes (as high and low) to see the general trend. Then, they were discretized into five classes, and the associations for the highest 20% biomass production, lipid content, and lipid production were also determined to identify the conditions leading to very high biomass production, lipid content, and lipid production. The results involving the top 20% performance were only used for overall evaluations because the number of data corresponding to this set was relatively small. For two-class (high and low) discretization, however, a more detailed analysis was conducted. The discretized output and descriptors are given in Table 4.2 and Table 4.3 respectively. It should be noted for the numbers in Table 4.2 and Table 4.3 that, brackets refer to lower or higher limits that are also included in the range (i.e. greater than or equal to), but regular parentheses refer to higher or lower limits that are not included in the range (i.e. greater than).

Table 4.2. Discretized classes of output variables.

	2-class discretization		**5-class discretization**	
Output variable	**low**	**high**	**very low**	**very high**
biomass production ($g.L^{-1}$)	[0, 11.2)	[11.2, 54.5)	[0, 6)	[18, 54.5]
lipid content (%)	[10, 22.8)	[22.8, 58.8)	[0, 10)	[35.5, 58.8]
lipid production ($g.L^{-1}$)	[0, 3.1)	[3.1, 19.4)	[0, 0.73)	[5.8, 19.4]

Table 4.3. Discretized classes of descriptors.

	level 1	**level 2**	**level 3**	**level 4**	**level 5**	**level 6**
C concentration ($g.L^{-1}$)	[0, 28)	[28, 56)	[56, 84)	[84, 112)	[112, 140)	[140, 168]
C/N molar ratio	[0, 56)	[56, 112)	[112, 168)	[168, 224)	[224, 283]	-
time (h)	[0, 30)	[30, 60)	[60, 90)	[90, 120)	[120, 180)	[180, 375]
mixing speed (rpm)	[0, 150]	(150, 200]	(200, 400]	(400, 800]	-	-
pH	[2, 6)	[6]	(6, 8]	-	-	-

4.2.2. Results of Decision Tree Analysis

To ensure that high and low productivities to be equally represented in the data, the median biomass production (11.2 g/L) was chosen as the split point for the decision tree. The tree was grown and optimized using 2/3 of the total data with the options described in Section 4.2. According to Bayesian optimization of the hyperparameters, the maximum number of splits was found as 107, minimum leaf size was discovered as 4 and the optimal split criterion was determined as “maximum deviance reduction”. After finding the optimal hyperparameters leading to the optimal tree structure, the remaining 1/3 of the data was used as a test set to see if the tree can classify data that has never been seen before.

Table 4.4 contains the confusion matrix displaying the classification accuracies of the classes. As it is indicated in the table, the tree classifies both high and low biomass productivities quite well as can be observed through the recall values shown in the last column for the training dataset. Additionally, the testing accuracy of the high values of biomass production is superior because of the recall value of 0.92. Similarly, when precision values are examined, both low and high productivities are well classified.

The optimal tree structure shown in Figure 4.3 has only 6% training error (meaning that only 6% of the biomass production is wrongly classified during the construction of the tree) and 12.9% testing error (meaning that the tree classifies only 12.9% of the data that has never been seen before). The tree has 8 alternative paths (green or yellow colored) leading to high biomass production, and those terminal nodes are numbered in parenthesis from left to right. The leaves with at least six high output values are considered generalizable, as indicated by the green color. The yellow paths also lead to high output values, but the terminal nodes reached this time do not have enough data points to be considered as generalizable.

The first of those paths (high 1), requires the initial C/N molar ratio to be small (<76), yeast extract and the $(NH_4)_2SO_4$ concentrations to be limited (<1.3g/L and <1.4g/L respectively), and the growth time to be greater than 14h. The second generalizable path (high 2) can be followed through the same branches (initial C/N molar ratio<76, yeast extract

concentration<1.3g/L, time≥14h) but this time requires $(NH_4)_2SO_4$ concentration to be above average (≥1.4g/L,) as well as the oxygen concentration to be high. Likewise, the third (high 3) and fourth (high 4) paths contain the same rules (initial C/N molar ratio<76, yeast extract concentration<1.3g/L, $(NH_4)_2SO_4$ concentration ≥1.4g/L, sufficient oxygen concentration, time≥57h); however, if the mixing speed is less than 260 rpm, the glucose concentration must be high (at least 20g/L); and if the mixing speed is greater than 260 rpm, no constraint related to glucose is required.

Table 4.4. Confusion matrix for biomass production.

	actual data		**estimations**		**classification**
	class	**#of data**	**low**	**high**	**accuracy (recall)**
Training	**low**	109	104	5	104/109=0.95
	high	124	9	115	115/124=0.93
	precision		104/113=0.92	115/120=0.96	
Testing	**low**	64	53	11	53/64=0.83
	high	52	4	48	48/52=0.92
	precision		53/57=0.93	48/59=0.81	

The rules related to the rest of the generalizable paths are: (high 5) initial C/N molar ratio<76, yeast extract concentration≥1.3g/L and time≥112h; (high 8) initial C/N molar ratio≥76, and $(NH_4)_2SO_4$ concentration ≥1.7g/L; (high 9) initial C/N molar ratio≥76, $(NH_4)_2SO_4$ concentration <1.7g/L, yeast extract concentration≥0.7g/L and yeast strain is SKY7.

Next, Figure 4.4 shows the descriptor importance analysis based on the usage frequency of the variables and the placement of the divisions during the construction of the tree. Before digging deeper into the specific details of this approach, some clarification on the meaning and potential benefits of this type of analysis is required. First and foremost, the findings of such an analysis are empirical and heavily dependent on the dataset used. If an important descriptor (such as oxygen sufficiency) has already been optimized by the research community, it may not be as significant as it should be in this analysis because almost everyone works in the already optimized range of this variable. In the light of this

information, fermentation time and C/N molar ratio appear to be more deterministic in this situation, followed by variables related to the nitrogen source.

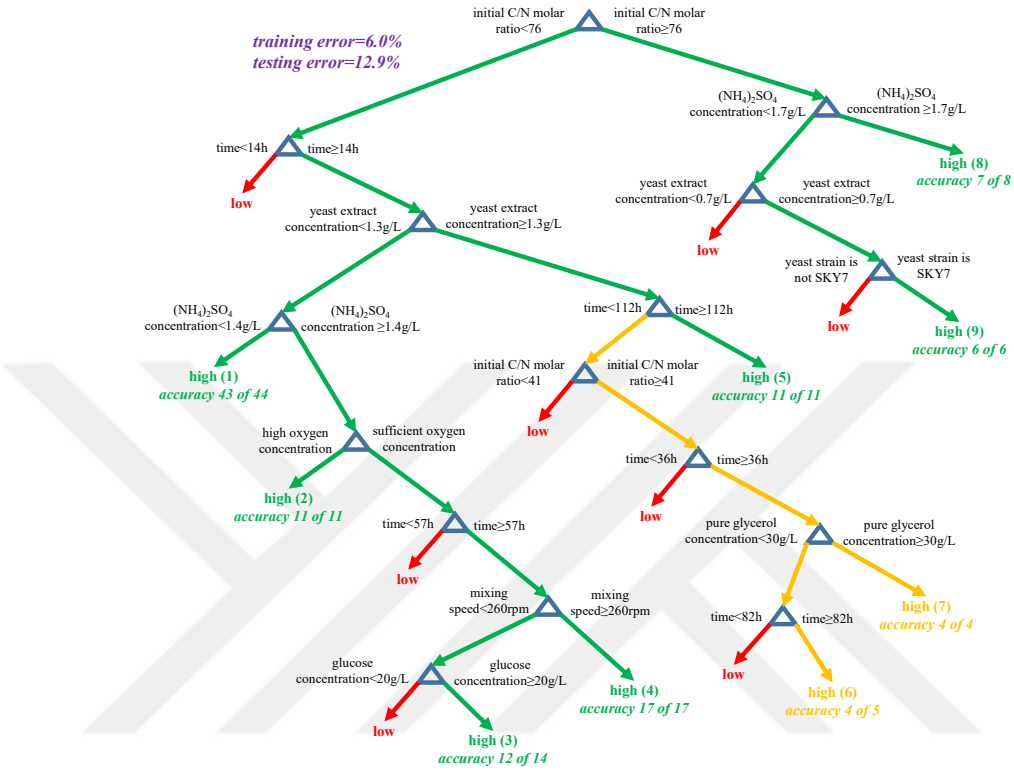

Figure 4.3. Decision tree classification for biomass production (green colored paths are generalizable) [154].

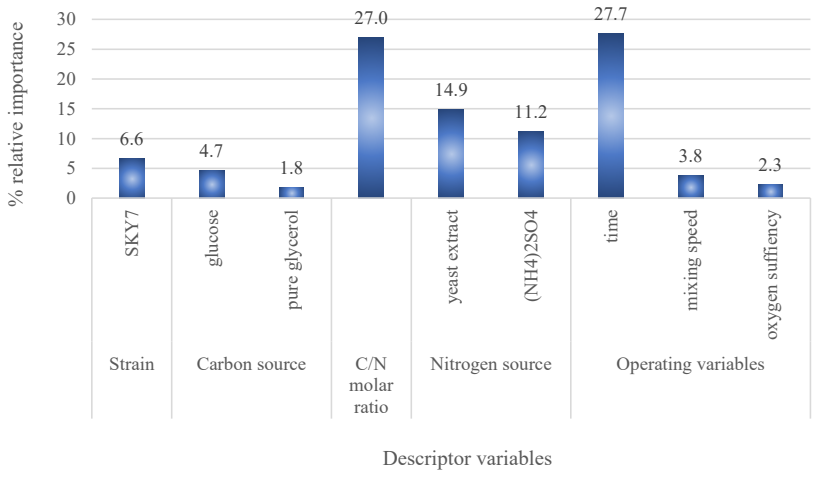

Figure 4.4. Relative percent importance of the descriptor variables on biomass production [154].

The next step is to employ a decision tree analysis to determine the pathways leading to high lipid content with a similar computational approach. Like the previous case, the median lipid content (22.8%) was chosen as the split point for the decision tree to ensure that high and low lipid contents were represented equally in the data. The maximum number of splits was found as 28, minimum leaf size was found as 2 and the optimal split criterion was determined as “Gini's diversity index”. The confusion matrix (Table 4.5) indicates that the accuracy and the precision of the model are quite satisfactory both for the training and the testing data.

Figure 4.5 shows the best decision tree structure, which has a training error of 7.3% and a testing error of 13.8. The three shows three generalizable paths leading to high lipid percentage (green colored branches). The first rule (high 3) is to keep the pH level smaller than 6.3, the initial C/N molar ratio smaller than 62, the glucose concentration to be high (≥5g/L) and the yeast strain not to be W29. The next rule (high 7) requires the pH level to be at least 6.3, crude glycerol concentration to be high (≥13g/L) and the time to be less than 89h. The final path (high 8) contains the same rules about the pH (≥6.3) and crude glycerol concentration (≥13g/L) but requires the yeast strain SKY7 to be grown at least for 89h.

Table 4.5. Confusion matrix for lipid percentage.

	actual data		**estimations**		**classification accuracy (recall)**
	class	**#of data**	**low**	**high**	
Training	**low**	110	105	5	105/110=0.95
	high	123	12	111	111/123=0.90
	precision		105/117=0.90	111/116=0.96	
Testing	**low**	68	55	13	55/68=0.81
	high	55	4	51	51/55=0.93
	precision		55/59=0.93	51/64=0.80	

The input importance analysis is shown in Figure 4.6. As it is indicated in the figure, pH level is the variable with the highest significance followed by glucose concentration and C/N molar ratio. Interestingly, the strain names are found to be deterministic to reach high lipid percentage as well.

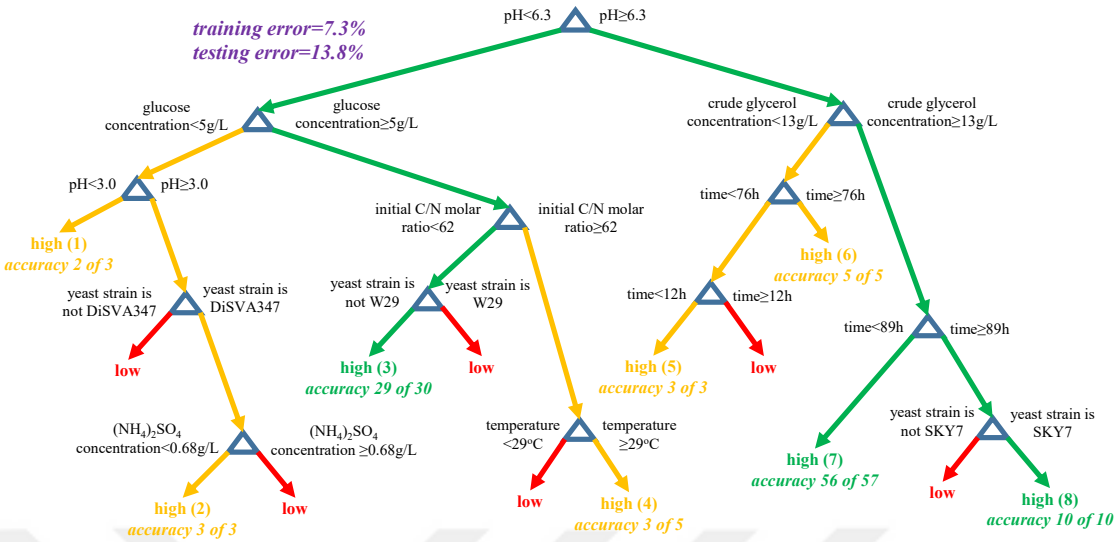

Figure 4.5. Decision tree classification for lipid percentage (green colored paths are generalizable) [154].

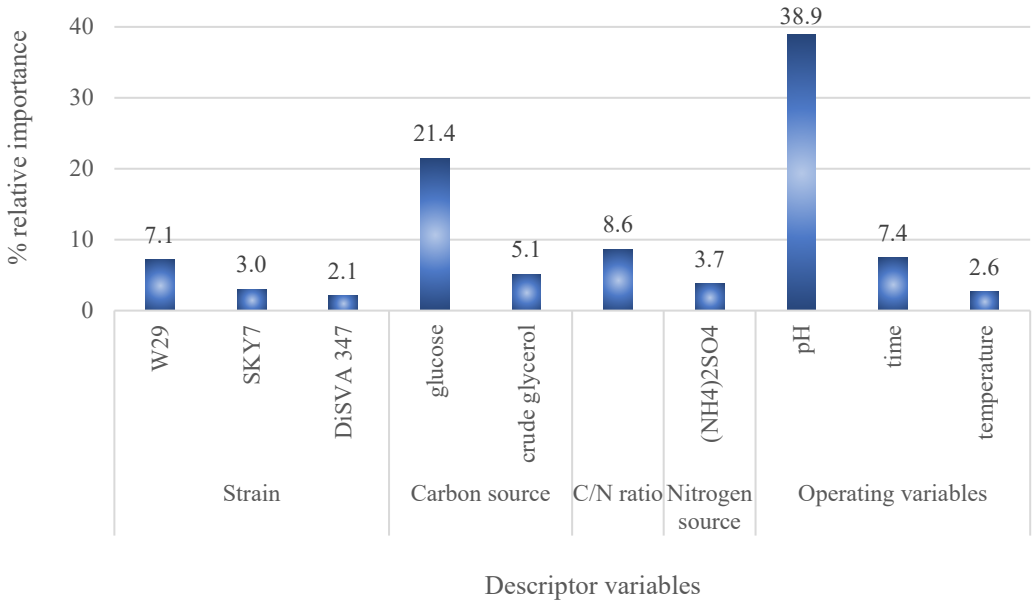

Figure 4.6. Relative percent importance of the descriptor variables on lipid percentage [154].

It should be noted that reaching higher lipid production should be the main goal to be able to produce biofuel that could compete with petroleum-based fuels. Hence, as the final step of this section, another decision tree was constructed (maximum number of splits=37,

minimum leaf size=1, optimal split criterion="Gini's diversity index") to classify the lipid production. The median lipid production (3.1 g/L) was used as the split point for the decision tree. The confusion matrix (Table 4.6) shows that the model's accuracy and precision for both training and testing data are quite satisfactory; however, the classification accuracies for high levels of lipid production are almost excellent (0.98 for the training set and 0.92 for the test set).

The optimal decision tree structure (training error=5.4%, testing error=12.5%) in Figure 4.7 indicates 5 generalizable paths leading to high lipid production. Among these, the path leading to "high 2" has a perfect accuracy (30 out of 30). The rule associated with this path requires pH to be less than 6.3, high glucose and $(NH_4)_2SO_4$ concentrations (≥5g/L and ≥0.9g/L respectively), lower concentration for yeast extract (<1.3g/L) and the yeast strain not be H-222. The next path having a large number of data points and a perfect accuracy is the one named as "high 3". This path needs a pH greater than 6.3, a C/N ratio smaller than 75, a time greater than 14 hours, and a strain other than W29. The other branches all come from this main branch but with a C/N ratio of more than 75. In these subbranches (leading to "high 4, "high 5" and "high 6"), the classification of biomass production is influenced by nitrogen and carbon sources, as well as oxygen level and time.

Table 4.6. Confusion matrix for lipid production.

	actual data		**estimations**		**classification accuracy (recall)**
	class	**#of data**	**low**	**high**	
Training	**low**	116	106	10	106/116=0.91
	high	107	2	105	105/107=0.98
	precision		106/108=0.98	105/115=0.91	
Testing	**low**	52	43	9	43/52=0.83
	high	60	5	55	55/60=0.92
	precision		43/48=0.90	55/64=0.86	

Finally, the input importance analysis (Figure 4.8) shows that pH level is the most significant variable, followed by $((NH_4)_2SO_4$ concentration and yeast extract concentration. Lipid production is also found to be affected by a variety of carbon sources and operating variables.

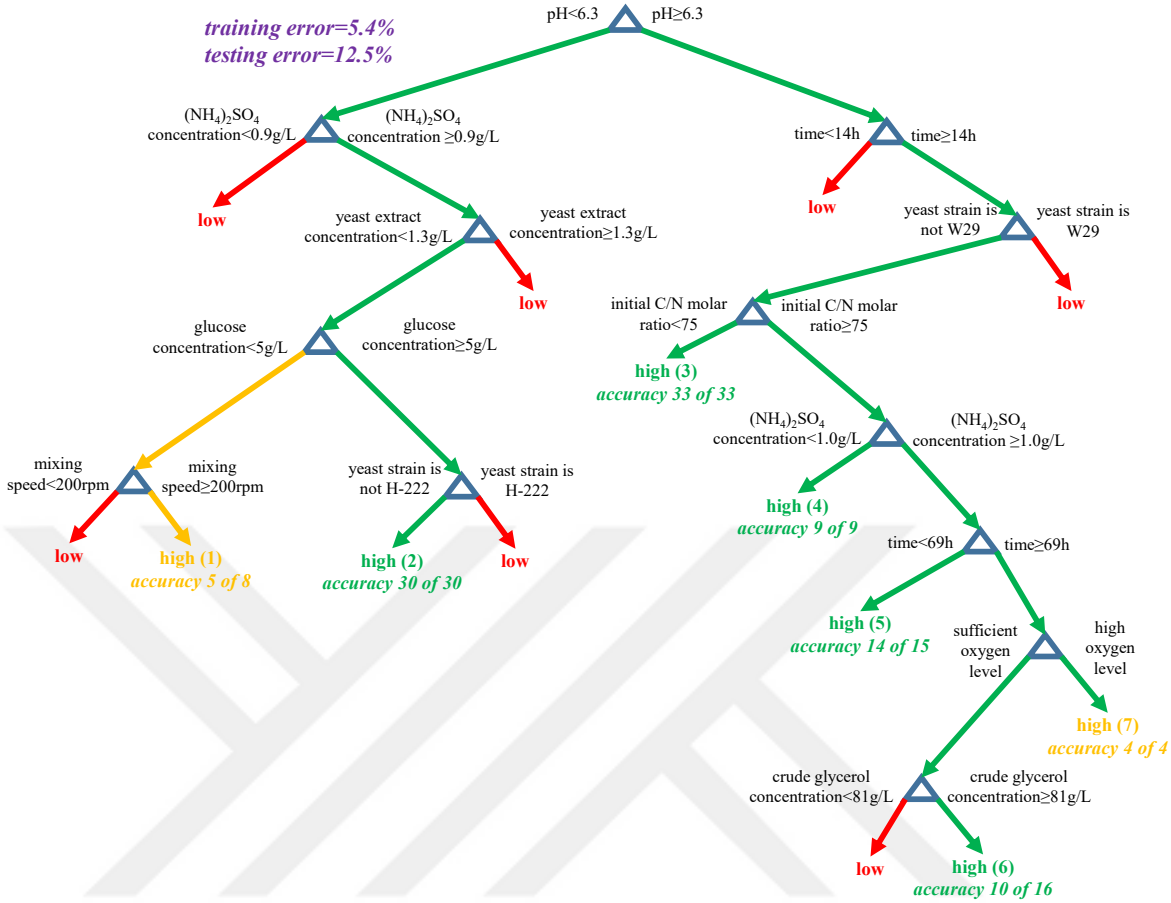

Figure 4.7. Decision tree classification for lipid production (green colored paths are generalizable) [154].

The most observable conclusion was that the statistical fitness's of DT models were quite high indicating the reliability of the results and rules deduced. Apparently, the use of only two levels as performance measures (low and high) played a significant role in this; using more than two levels resulted in lower statistical fitness's, forcing us to prefer rougher but more reliable results over finer but less reliable conclusions. With this respect, the decision tree was quite successful to determine the combination of features leading to high biomass production, lipid content and lipid production. As an example, 43 out of 44 cases resulted in high biomass production when the initial C/N molar ratio was low (<76), yeast extract and the $(NH_4)_2SO_4$ concentrations were limited (<1.3g/L and <1.4g/L respectively), and the growth time was greater than 14h. Similarly, 29 out of 30 cases involving pH levels smaller than 6.3, the initial C/N molar ratio smaller than 62, high (≥5g/L) glucose concentration led to high lipid content if other than W29 was used as the yeast strain. The

C/N ratio and fermentation time were found to be the most influential features for biomass production while the use of glucose and medium pH seemed to be more important for high lipid content. For the lipid production case, five generalizable paths leading to high values of this output were identified. One of those paths required pH to be less than 6.3, high glucose and $(NH_4)_2SO_4$ concentrations (≥5g/L and ≥0.9g/L respectively), lower concentration for yeast extract (<1.3g/L) and the yeast strain not to be H-222. Another one needed a pH greater than 6.3, a C/N ratio smaller than 75, a time greater than 14 hours, and a strain other than W29.

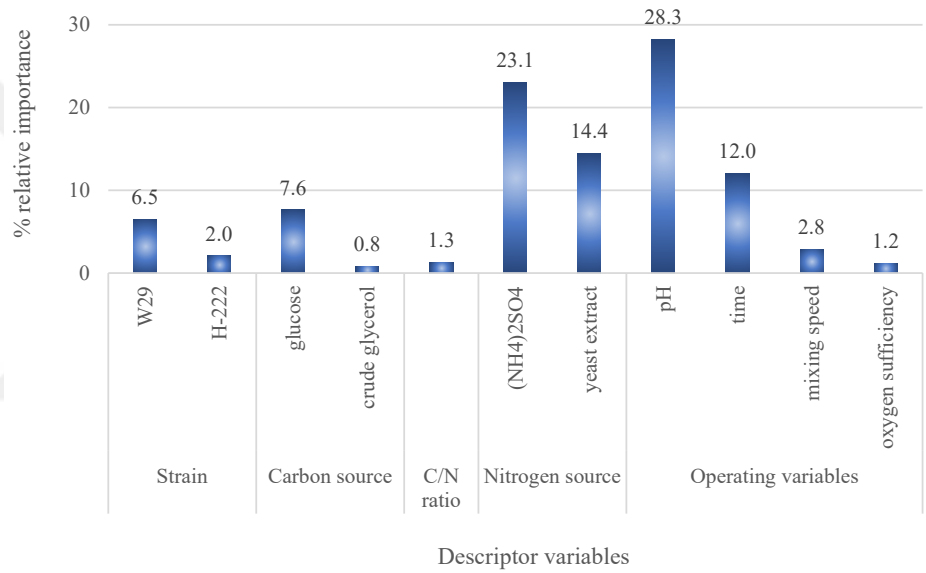

Figure 4.8. Relative percent importance of the descriptor variables on lipid production [154].

4.2.3. Results of Association Rule Mining

The association rule mining algorithm was used to divide the output variables chosen for this study (biomass production, lipid content, and lipid production) into two and five equal-sized classes as shown in Table 4.7. It was aimed to determine the effects of individual descriptors and support the results of the decision tree analysis presented above. First, the 2-class discretization was used to deduce general rules to explore the factors that lead to greater than the median of the particular output (top half of data). Then, by limiting the concept of high production to the top 20% of the data by 5-class discretization, more strict rules were

gathered compared to 2-class discretization. Although 5-class discretization could be solely used to gather information yielding very high output class, the limited number of data, in this case, could be misleading for the analysis. Consequently, association rule mining analysis was done for 5-class just for general evaluation while the associations obtained by 2-class discretization were discussed in detail to understand the performance at different levels of descriptors. Table 4.7 shows some of the most important results (based on lift) for both high (top 50%) and very high (top 20%) biomass production and lipid content.

To measure the strength of associations, the table includes three parameters: support, confidence, and lift. The fraction of a range or class (for example glucose) of a feature (for example carbon source) that yields the corresponding class of the output variable (for example high biomass production) in the total database is referred to as “support”. On the other hand, the ratio of the corresponding class of the output variable at a level of an input variable (i.e. high biomass production data using glucose) is referred to as "confidence". For example, the first row of Table 4.7 shows the association rule parameters for glucose as a carbon source. There are 60 high biomass production cases obtained with glucose as the carbon source. The total number of cases with glucose is 83 and the total number of cases in the database is 356. Consequently, the support value for this rule is calculated as 60/356=0.17 while the confidence value is 60/83=0.72, meaning that, 72% of the cases with glucose leads to high biomass production. Finally, the last parameter for association rule mining is the lift, which is the main indicator for the positive or negative effect of a level of an input variable providing very high biomass production. It is calculated by dividing the confidence of the rule by the fraction of the particular level of the output (i.e. total high-class data) in the total database. For the case of using glucose as the carbon source, the lift is calculated as 0.72/(179/356)=1.43. This number indicates that the fraction of high biomass production among the cases using glucose is 1.43 times higher than the fraction of high biomass production cases in the entire dataset. This is clear evidence that glucose is superior to other carbon sources in terms of high biomass production. The other rules in Table 4.7 can also be interpreted in the same way: the higher the lift value from unity, the stronger the link between that factor and high biomass production.

As it is described above, glucose as the carbon source is important for the high biomass production (lift is 1.43); however, the lift decreases when the biomass growth requirement

was tightened to the top 20% indicating that some other factors become more important for the very high biomass production. On the other hand, glucose, which is also highly associated with high lipid content (top 50%), becomes more influential for very high lipid content (top 20%) considering that the lift increased from 1.3 to 1.79. As expected, the fermentation time of 30-60 h is associated with high (lift is 1.33) and very high (lift is 1.87) biomass production. Higher fermentation times (90-120 h), on the other hand, promote very high lipid content (lift of 1.71 compared to 1.12 for high class). Finally, pH higher than 6 seems to be good for high biomass production and lipid content.

Table 4.7. Selected association rule mining results.

Input Variable *(feature A)*	**Output** *(biomass production ($g.L^{-1}$) OR lipid content (%))*	**Output Class** *(class B)*	**Count** *(number of cases satisfying this condition)*	**Support** $P(A \cap B)$ *(fraction of cases with feature A and class B to the total number of cases in the dataset)*	**Confidence** $\frac{P(A \cap B)}{P(A)}$ *(fraction of cases with feature A and class B to the total number of cases with feature A)*	**Lift** $\left(\frac{P(A \cap B)}{P(A)}\right) \Big/ P(\mathrm{B})$ *(fraction of "confidence of the rule" to "ratio of cases of class B in the total number of cases")*
Carbon source = glucose	biomass production	high	60	0.17	0.72	1.43
		very high	13	0.04	0.16	0.75
	lipid content	high	54	0.15	0.65	1.3
		very high	30	0.08	0.36	1.79
	lipid production	high	53	0.16	0.64	1.29
		very high	13	0.04	0.16	0.79
Time = 30-60 h	biomass production	high	43	0.12	0.67	1.33
		very high	25	0.07	0.39	1.87
Time = 90-120 h	lipid content	high	47	0.13	0.56	1.12
		very high	29	0.08	0.36	1.71
Time = 30-60 h	lipid production	high	44	0.13	0.69	1.39
		very high	17	0.05	0.27	1.33
pH = >6	biomass production	high	92	0.26	0.64	1.28
		very high	55	0.16	0.38	1.84
	lipid content	high	111	0.31	0.80	1.60
		very high	44	0.12	0.32	1.57
	lipid production	high	110	0.33	0.79	1.60
		very high	54	0.16	0.39	1.94

The lift values for a variety of association rules derived for the high (top 50%) biomass production, and lipid content at different levels of input variables are shown in the following figures. Hence, the change of lifts as a function of different levels of input variables can be observed better to deduce more detailed information about the effect of specific input variables. The blue circles represent the top 50% of the output class while the orange circles indicate the bottom 50% of the output. The sizes of the circles are proportional to the number of cases, which are also given as numbers in the circles. For instance, Figure 4.9 shows lift values for the association rules for different strains of *Yarrowia lipolytica*. The strains SKY7 and SM7 are good for both high biomass production and lipid content as evident from the higher lift of high class and the large number of cases supporting this result (for higher reliability) while W29 seems to be just the opposite. Although there are also strains that produced either purely high or low performances, they are extracted from single articles and not verified by the other sources in the dataset (circled with black borders).

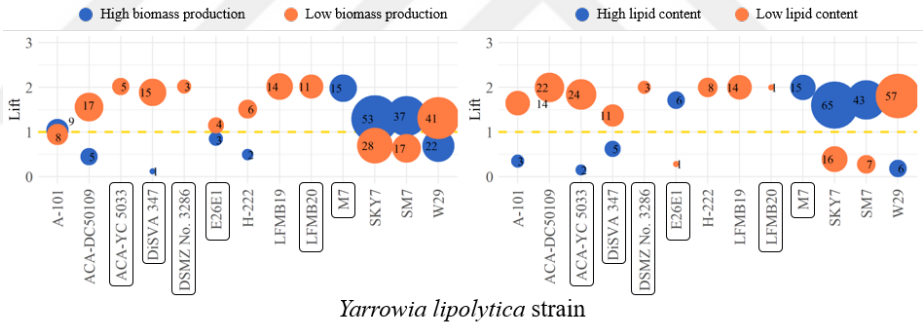

Figure 4.9. Lift and count values for the association rules derived for different strains of *Yarrowia lipolytica* (data related to the strains circled by black color were extracted from single articles and not verified by the other sources in the dataset) [154].

Aside from the strain type, Figure 4.10 and Figure 4.11 depict the relationship between biomass production and lipid content as a function of culture conditions (reactor type, fermentation time, pH value, and fermenter mixing speed) and nutrients (carbon source, carbon level, C/N ratio, and oxygen supply). Indeed, there are also other important factors that may affect the lipid production of *Y. lipolytica* such as magnesium [348], extraction method [356, 360], temperature [342], nitrogen source [361], culture volume [346], minerals (sulfur, zinc, phosphorus, etc.) [362] and also vitamins, like biotin [363]. However, no

detailed analysis for those variables could be performed because most of the experimental studies in database were conducted under similar conditions in terms of these variables. As shown in Figure 4.10a, fed-batch reactors favor high biomass production. In the study of Chen *et al.* [364] it was also concluded that fed-batch reactors yield higher biomass production than batch reactors for oleaginous yeasts. Although batch reactors appear to perform better in terms of lipid content, no significant rule can be deduced from the dataset.

The growth cycle of *Y. lipolytica* is very much time-dependent and distinct phases can be identified; such as biomass growth phase, lipogenic phase and citric acid production phase [365]. Liu *et al.* [366] concluded that *Y. lipolytica* exhibits two phases, where the growth phase is in between 24 and 44 hours when nitrogen is present in the medium, and a following lipid production phase when nitrogen is depleted and available carbon is stored as lipid between 56 and 76 hours. In another study, it was found that 72 hours is the optimum fermentation time for lipid production of *Y. lipolytica* [342]. Moreover, Dourou *et al.* [367] identified the phases of *Y. lipolytica* ACA DC 50109 as balanced growth until 24 hours, oleaginous phase (24-120 hours), and finally lipid turnover phase with the depletion of carbon source. From association rule mining results in Figure 4.10b, it is found that high biomass production is mostly favored in between 30 and 60 hours. From Figure 4.10b it is also concluded that lipid content decreases after 120 hours, with a maximum value between 60 and 90 hours.

The optimal pH value for lipid production from *Y. lipolytica* varies with strain and carbon source [368]. Gao *et al.* [369] also found that an initial pH of 8 is regarded as the optimum value for different carbon concentrations for *Y. lipolytica* CICC 31596. From association rule mining results shown in Figure 4.10c, however, it is found that initial pH greater than 6 generally favors both high biomass and lipid content.

Agitation rate is closely related to the oxygen level in the growth medium; such that, as the agitation rate increases the amount of oxygen that dissolves in the medium increases. Enshaeieh *et al.* [342] experimented on the effect of agitation speed on lipid production and found that the optimum agitation speed is around 200 rpm. Unrean *et al.* [143] also reported that lowering the agitation speed in batch shake flasks from 300 to 100 rpm leads to an

increase in lipid accumulation. According to the results of association rule mining (Figure 4.10d), high agitation speed favors high biomass production as expected. However, the highest lift value for lipid content is obtained between 200 and 400 rpm levels.

Figure 4.10. Lift and count values for the association rules derived for different culture conditions of *Yarrowia lipolytica*: (a) reactor type, (b) fermentation time, (c) pH, (d) agitation rate (rpm) [154].

It should be noted by observing Figure 4.10 (and the remaining figures) that the lift for the high and low biomass and lipid production are close to each other (and therefore to one).

This behavior is particularly dominant in the medium values of descriptors as expected because the intermediate values of descriptors may lead to high or low levels of performance depending on the level of other descriptors. Another plausible reason for such behavior is that the range of related descriptors is well established/optimized, and everyone tends to use those descriptors within this range. In such cases, the use of the lower or higher values (within that range) of the descriptor cannot make significant differences in the outcome.

Different kinds of carbon sources were experimentally tested for the growth of *Y. lipolytica* from conventional glycerol and glucose to unconventional sources like acetate [366] and hydrolysate from subcritical water treatment of Jatropha curcas L. seed kernels [370] in the literature. Different hydrophilic (glucose, glycerol, and alcohols, etc.) and hydrophobic substrates (fatty acids, triacylglycerols, and alkanes, etc.) can be used as carbon sources for *Y. lipolytica* [368]. From association rule mining results shown in Figure 4.11a, it can be deduced that glucose is the best option that favors high lipid content with high biomass production. Some studies concluded that crude glycerol generated by biodiesel synthesis can be used as the sole carbon source for *Y. lipolytica* production without any significant change in lipid content [345, 371]. This is important in terms of biorefinery engineering strategy that uses the by-product waste of one process as a nutrient for another process for improving cost-effectiveness [363]. According to the association rule mining results in Figure 4.11a, crude glycerol is also found not to affect lipid content either positively or negatively.

It was shown that that increasing carbon concentration increases the lipid content until a certain concentration and starts to decrease afterwards for *Yarrowia lipolytica* ACA-YC 5033 and *Yarrowia lipolytica* SKY7 [340, 349]. Bao *et al.* [368] also concluded that excessively low and high carbon concentrations have negative effects on biomass and lipid production. Chang *et al.* [372] discussed the inhibitory effect of glucose concentration, stating that when glucose concentration in the growth media exceeds a certain weight percent lipid content of *Cryptococcus* sp. starts to decrease. According to the results of association rule mining shown in Figure 4.11b, increasing carbon level has a positive effect on biomass production and lipid content until a certain range (75-100 g/L for biomass production and 75-100 g/L for lipid content), after which it favors low production.

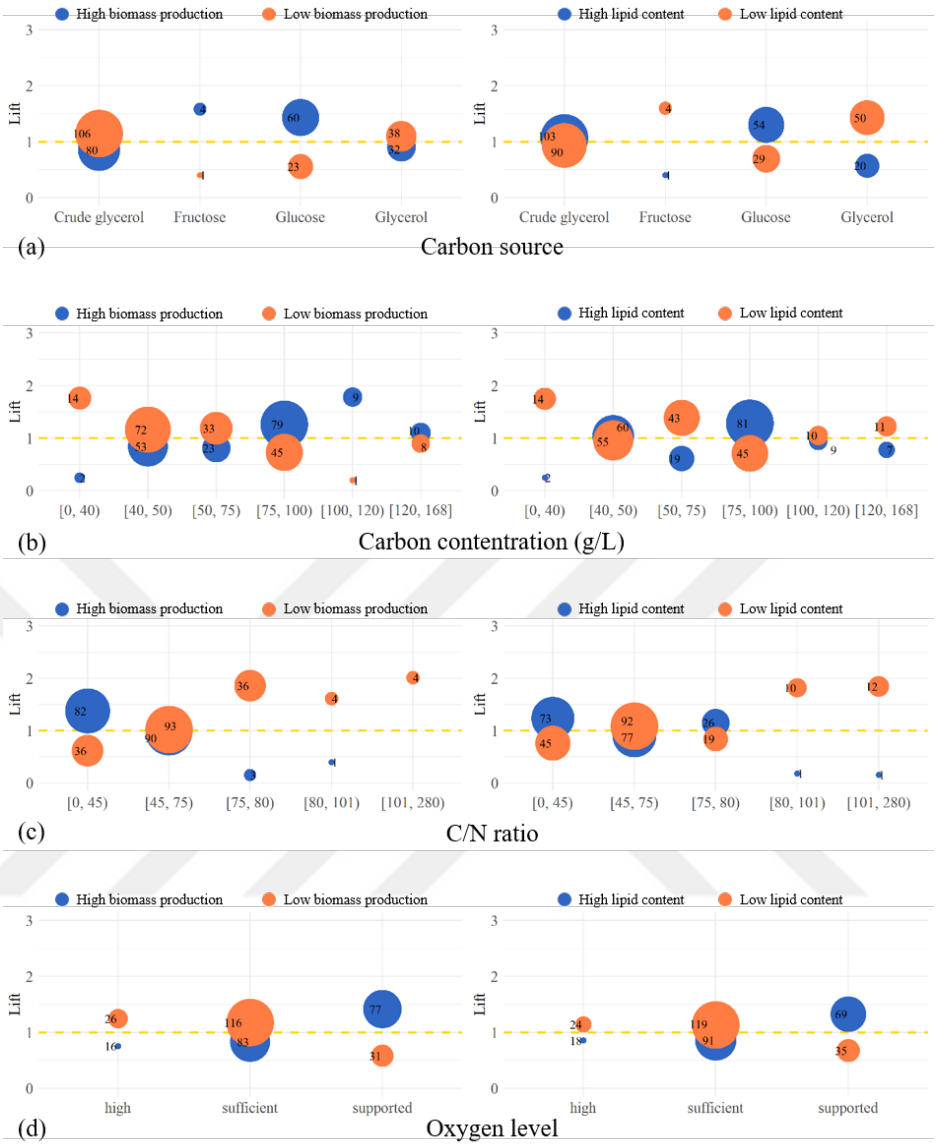

Figure 4.11. Lift and count values for the association rules derived for nutrient elements of *Y. lipolytica*: (a) carbon source, (b) carbon concentration (g/L), (c) C/N ratio, (d) oxygen level [154].

It was reported that in oleaginous yeasts, the driving force for lipid accumulation is the nutrient imbalance between excess carbon and limited nitrogen concentration, however, *Yarrowia lipolytica* may also favor citric acid production over lipid [373, 374]. C/N ratio is an important parameter for the activation of lipid and citric acid production pathways in *Yarrowia lipolytica* [374]. A higher C/N ratio seems to induce the accumulation of organic acid byproducts instead of lipids [375]. The results of ARM on lipid content in Figure 4.11c

also show a mixed pattern, with a lower and medium C/N ratio promoting high lipid content and a higher C/N ratio promoting lower lipid content. Although the number of data is very limited and no concrete reasoning can be made, this may be the result of the citric acid production over lipid. In a previous work, it was concluded that C/N ratios ranging from 40 to 80 are the preferred range for most of the oleaginous microorganisms for lipid accumulation [368]. Kolouchova *et al.* [361] also found that a high C/N ratio leads to high lipid production, but that when the ratio exceeds 100, the behavior reverses, resulting in a decrease in lipid production. According to a review article, lipid accumulation is stimulated when the C/N ratio is greater than 20; however, when the C/N ratio is greater than 70, lipid production can decrease in some cases [362].

The experiments on oxygen availability in the field show mixed behavior. Although in some studies it was concluded that increasing the oxygen availability in the culture medium increases the lipid content [376], in some other works it was found that higher aeration decreases the lipid accumulation in highly aerated environments [377]. Indeed, the importance of oxygen availability on *Yarrowia lipolytica* in lipid biosynthesis is discussed in a recent review article [375]. According to association rule mining results shown in Figure 4.11d, it was found that increased oxygen availability favors high biomass production and lipid content.

ARM results for lipid production of *Yarrowia lipolytica* are shown in Figure 4.12. Strain M7 shows the highest lift value for high lipid production as shown in Figure 4.12a, which has also the highest lift value in high biomass production and high lipid content as shown in Figure 4.9. In Figure 4.12b, reactor types are plotted against the lift values, and the fed-batch reactor is found to have the highest lift value for high lipid production. Although batch reactor is found to have the highest lift value in high lipid content, high biomass production in fed-batch reactors resulted in the highest lift value in lipid production. Lift values of lipid production for fermentation time (Figure 4.12c) and pH (Figure 4.12d) show similar behavior with lipid content while favoring the highest lift value even more. Similarly, lift values of lipid production for fermenter mixing speed (Figure 4.12e), carbon concentration (Figure 4.12g), C/N ratio (Figure 12h), and oxygen level (Figure 4.12i) shows similar behavior with biomass production, while favoring the highest lift value even more.

In Figure 4.12f, lift values of lipid production for major carbon sources are shown. Glucose is found to be the optimum carbon source with also having the highest lift value in high biomass production and high lipid content, whereas fructose is found to have the highest lift value for low lipid production because of the mixed behavior in lipid content and biomass production as shown in Figure 4.11a.

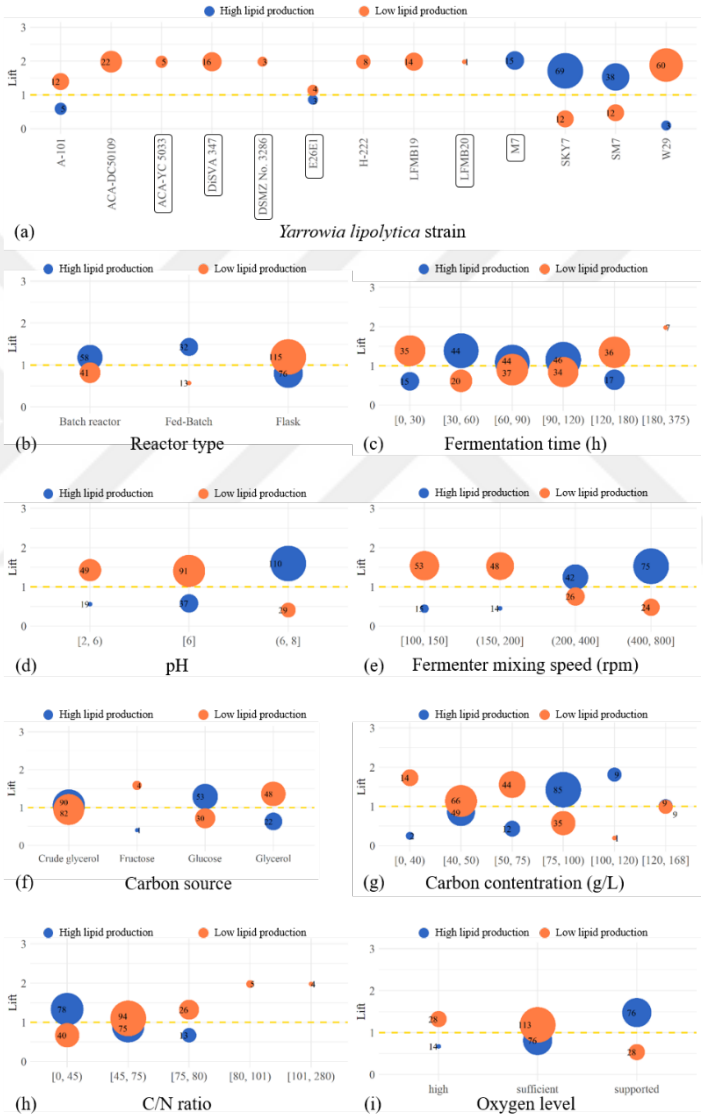

Figure 4.12. Lift and count values for the association rules derived for lipid production of *Yarrowia lipolytica*: (a) strain, (b) reactor type, (c) fermentation time, (d) pH, (e) agitation rate (rpm), (f) carbon source, (g) carbon concentration (g/L), (h) C/N ratio, (i) oxygen level [154].

ARM provided more specific results involving the specific strains that were used, operational conditions and nutrients. *Yarrowia lipolytica* strain M7 was found to be the most promising strain for both biomass production and lipid production. Fermentation time was an important factor for yeast growth; 30-60 hours was found to be the optimum time for biomass production and lipid production, whereas 90-120 hours was determined to be the optimum for lipid content. Fed-batch operation was found to increase biomass and lipid production while glucose was found to be the optimum carbon source within the options available in the dataset for both high biomass production and high lipid content. Carbon concentration was also an important factor that needed to be optimized. Lastly, it was found that conditions with pH above 6 significantly favored high biomass production and lipid content and consequently lipid production as well.

To sum it up, although some of the deduced trends and rules were already known in the literature, methods studied here revealed some previously unknown rules and patterns, such as the effect of pH on lipid content and lipid production. Consequently, machine learning methods applied to the yeast database can reveal the best growth conditions for maximum lipid yield in biofuel manufacturing which can guide the planning and implementation of new experimental works.

5. ML ANALYSIS OF TRI-REFORMING OF METHANE

In this part of study, a database for methane tri-reforming (oxidative steam reforming of biogas) is compiled from literature data and analyzed using various machine learning tools to predict the CH_4 conversion, CO_2 conversion and syngas ratio in tri-reforming process. For this purpose, a comprehensive database was constructed by reviewing scientific papers on biogas tri-reforming published between 2004 and 2024, sourced from "Web of Science" database. After a detailed preprocessing, the dataset was finalized with 1183 data entries from 116 experiments across 29 articles published in the literature.

Input variables (descriptors) were chosen as the type of catalyst, catalyst preparation conditions and methods, reaction conditions and feed gas compositions while the output (target) variables were CH_4 conversion (%), CO_2 conversion (%) and H_2/CO ratio. Since the articles used in this study did not report all three target variables, the size of the data was different for each target. Accordingly, CH_4 conversion, CO_2 conversion, and H_2/CO molar ratio datasets have 938, 693, and 644 data entries respectively. The descriptors with their ranges and identities are presented in Table 5.1 while the distribution of target variables is shown in Figure 5.1.

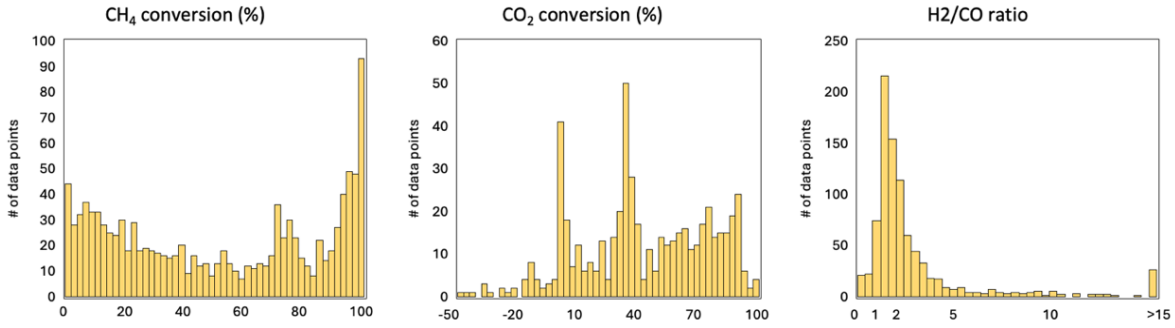

Figure 5.1. Frequency distribution of data points over the (a) CH_4 conversion, (b) CO_2 conversion, (c) H_2/CO conversion range.

Studies included into dataset is limited to the data only obtained in the flow reactor using particulate catalysts, while some studies were excluded because they require the use

of different descriptor sets and/or they may have conditions that may not repeated in frequencies that are not sufficient for effective training of the models. The examples are; (i) studies over pellet catalysts [378-380], (ii) studies on the effect of impurities [157, 381-383], (iii) ionic liquid-assisted catalyst preparation [384, 385], (iv) reactions in electric field [386], (v) works involved in studying catalyst preparation conditions such as, calcination gas, precursor preparation method, precursor type [387, 388], (vi) perovskite type catalysts [389, 390], (vii) pressure other than 1 atm [391]. Some operational variables are also excluded from the dataset due to limited reporting in the articles. These include some catalyst characteristics (e.g., surface area, pore volume, average pore diameter), certain catalyst preparation related variables (e.g., dehydration temp.), and some of the reaction conditions (e.g., reactor type, pressure, catalyst amount).

Table 5.1. Input variables and their ranges.

Input variable	**Ranges for continuous variables or identities for discrete variables**
metal	Ni, Co, Pt, Rh, Mo, Fe, Ru, Cu
support type	ZnO, La_2O_3, Al_2O_3, α-Al_2O_3, ɣ-Al_2O_3, $MgAl_2O_4$, MgO, CeO_2, ZrO_2, SrO_2, β-SiC, cordierite, SBA-15, Y_2O_3, KNO_3, Sm_2O_3, Nb_2O_5, MWCNT, TiO_2, hydrotalcite (Mg-Al-C), SiO_2
calcination time (h)	2-6
calcination temperature (°C)	350-1350
H_2% in reduction medium	4-100
reduction time (h)	0.5-4
reduction temperature (°C)	400-960
reaction temperature (°C)	407-950
time on stream (TOS) (min)	maximum 7200
W/F	
CH_4% in reaction stream	5.3-55.5
CO_2% in reaction stream	2.6-48.3
H_2O% in reaction stream	0.4-61.0
O_2% in reaction stream	0.6-28.9
Ar, He or N_2 % in reaction stream	0-87

The study started with the literature review to analyze the effects of descriptors on output variables in Section 5.1. Then, in Section 5.2, ML models are developed with predictive artificial neural networks (ANN), and random forest (RF) to predict the outcome

of untested conditions, and decision tree classification (DT) is used to identify possible heuristic rules to determine the conditions leading to the desired levels of the target variables. Following these ML methods, in Section 5.3, Transfer Learning methodology is implemented on ANN models to explore the capability of this methodology to increase the predictive accuracy of ANNs. Described dataset is used in in the analysis explained in Section 5.2 and Section 5.3.

5.1. Pre-analysis of the Variables Affecting Tri-reforming Process

In Figure 5.2, the number of publications that reported the effect of individual descriptors is given. The most frequently studied descriptors are the feed composition, reaction temperature, time on stream (TOS), and gas hourly space velocity (GHSV). We will briefly summarize the findings related to the effects of these descriptors on catalytic performance before discussing the machine learning model.

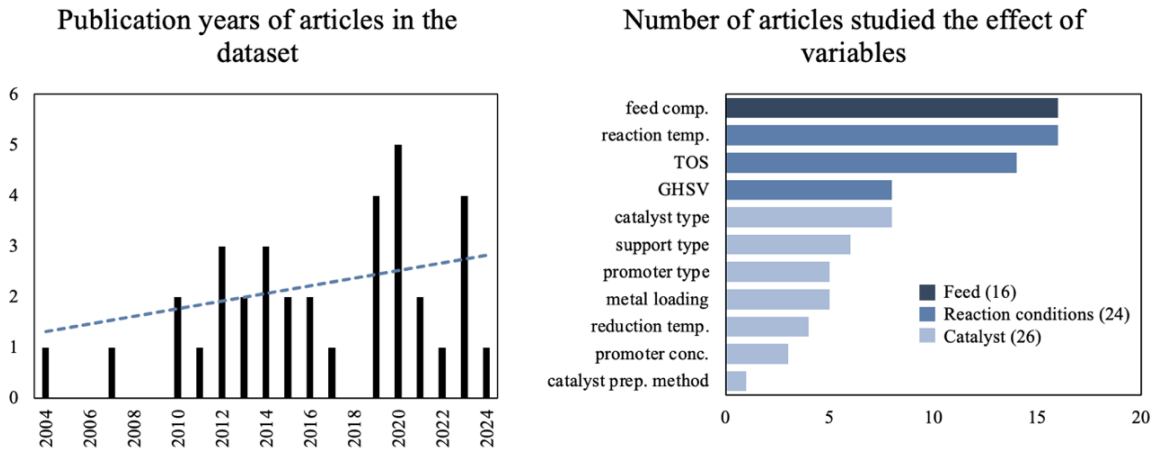

Figure 5.2. (a) Number of publications over the years in the dataset. The dotted lines are the least-squares trend lines (b) Number of publications studied the effect of different variables.

As different feed components favor different reactions, the feed composition has a significant influence on the product distribution. Hence, the feed composition, especially the effects of various feed components was investigated extensively.

Various investigators [162, 392, 393] observed that the CH_4 conversion generally increased with increasing O_2 concentration in the feed. In contrast, the increase in O_2 in the reaction medium negatively affects the CO_2 conversion [394-396]. It is also observed by various researchers that the O_2 level in the reaction medium had little effect on the H_2/CO ratio within the investigated range [393, 396, 397]. Pino *et al.* [398], on the other hand, reported a slight increase in H_2/CO ratio from 1.62 to 1.70 with higher O_2, alongside a notable reduction in the heat of reaction. Similarly, Kumar *et al.* [399] found that an increase in the O_2 mole ratio from 0.1 to 0.4 results in an increased H_2/CO ratio while Lino *et al.* [394] stated the increase in the O_2 concentration usually favors POM over DRM, resulting in lower CO_2 conversion and producing syngas with higher H_2 content. Oxygen has a high affinity towards methane; high O_2 concentration enhances methane oxidation and limits dry reforming as the reaction of methane with oxygen is thermodynamically favored over CO_2. Oxygen was also found to be consumed faster than other oxidants [400]. Additionally, higher O_2 concentration leads to partial or total oxidation of CH_4 [401], converting it to CO_2 and H_2O, which are then reformed into CO and H_2 by DRM and SRM for improved syngas yield [399].

Another CO_2-forming reaction is the WGS reaction. At high O_2 content, CO_2 conversion is affected by CO_2 produced from oxidation of CH_4 and CO_2 consumed by dry reforming. The heat associated with CH_4 oxidation, especially at lower H_2O/CH_4 ratios, can impact the temperature gradient within the reactor, maintaining high reactor temperatures, limiting WGS reaction, and promoting RWGS. This helps to decrease H_2/CO molar ratio by increasing CO in the reaction products [396]. Oxidation of methane is highly exothermic, and increased O_2 concentration in the reaction medium release more energy. This heat supports endothermic reactions, reducing the external energy required to obtain higher CH_4 conversion [400]. On the other hand, CO_2 conversion may also result from the oxidation of carbon deposits formed from methane cracking or the Boudouard reaction [401]. Addition of O_2 leads to the oxidation of methane and CO_2 production, affecting net CO_2 conversion. CO_2 conversion could also be influenced by the coke oxidation reaction [397].

Research studies have consistently shown that the addition of steam into the TRM reaction has been shown to favor SRM while reducing the conversion of CO_2 [162, 395-397,

399, 400]. Other experimental studies indicate that increasing H_2O content in the feed enhances CH_4 conversion [392, 395-397, 399, 400]. Although the majority of the studies found that the addition of steam increases H_2/CO molar ratio, suggesting that a higher H_2O/CO_2 ratio favors the SRM and suppresses DRM [396, 399, 402], Singha *et al.* [400] found no change in H_2/CO ratio with increased H_2O concentration.

Studies suggests that steam competes with CO_2 for catalyst active sites, resulting in reduced CO_2 conversion. Thermodynamically, CH_4 prefers to react with H_2O over CO_2 leading to a reduction in DRM with increased steam [162, 397]. Comparatively lower endothermicity of SRM favors the reaction of steam compared to CO_2 at high temperatures [399]. Also, increased steam ratio in the feed promotes WGS reaction and increases H_2/CO molar ratio [162, 396]. Moreover, H_2O can be both a reactant and a product in the overall process, highlighting the importance of optimizing the feed ratio for maximum selectivity and desired product yield. Additionally, H_2O helps mitigate carbon deposition, which is a common issue in DRM [400].

A reduction in the amount of the main reactant in the feed stream increases the contact time of the reactants to the catalyst and the conversion rate increases. It was observed that reducing the amount of methane in the feed leads to higher methane conversion [392].

The effect of varying CO_2 ratios in the feed is not straightforward. In experiments with increased CO_2 mole ratio led to higher methane conversion at low steam content conditions [392, 399]. It was also noted that increasing the CO_2 mole ratio decreased CO_2 conversion and reduced the H_2/CO ratio [399]. This behavior is attributed to competition among CO_2, H_2O and O_2 with methane, where H_2O and O_2 react preferentially over CO_2, for the same active catalyst sites [392]. Experiments with a low CO_2 ratio in feed, influenced by the WGS reaction, resulted in the lowest CO_2 conversion [392]. The activity of WGS reaction becomes dominant at feed composition with high steam and low CO_2 and at low reaction temperatures, resulting in minimal CO_2 conversion and even net CO_2 production [392].

In summary, the findings from the articles can be generalized as follows; an increase in oxygen in the feed enhances methane conversion, reduces CO_2 conversion, and has no significant effect on the H_2/CO ratio. An increase in H_2O also increases methane conversion, decreases CO_2 conversion, and slightly increases H_2/CO ratio. Lower methane content in the feed results in higher methane conversion. The effect of varying CO_2 ratios in the feed shows mixed behavior, influenced by other factors such as H_2O and O_2 ratio and reaction temperature.

Carbon deposition on the catalyst surface from methane cracking and the Boudouard reaction and catalyst metal sintering are primary causes of catalyst deactivation in methane reforming [402]. One of the main advantages of TRM is its resistance to coke formation, mainly due to the presence of oxygen and steam in the feed, which enhances coke removal mechanisms. The stability of the catalysts for methane reforming depends on factors such as; metal support interaction, active metal particle size, catalyst morphology, and metal dispersion [400]. To illustrate the importance of these factors, it is discussed that high metal dispersion reduces active metal particle size, increasing the number of active sites and enhancing resistance to metal sintering [400]. Also, catalysts with high surface area, large pore volume, and large average pore diameter facilitate the internal diffusion of reactants and products [403].

Yoo *et al.* [403] tested high metal dispersion and strong metal support interaction catalyst, observing no noticeable deactivation after 600 min reaction. Kumar *et al.* [404] synthesized stable Ni catalysts with strong resistance to metal sintering and coking. Ni metal confinement in SBA-15 hexagonal pores suppressed carbon deposition and metal sintering. It is discussed that smaller Ni particle sizes were more resistant to carbon deposition. ZrO2 supported catalysts. On the other hand, retained Ni in a metallic state resulting from the oxophilic nature of ZrO_2, which prefers interaction with oxygen prevents re-oxidation of Ni. Although catalyst performance remained largely unaffected, Walker *et al.* [405] noted the importance of calcination temperature, as a decrease in catalyst surface area is observed due to higher reaction temperature compared to the calcination temperature. Pandey *et al.* [406] synthesized a stable catalyst with consistent performance over 50h and attributed this to the catalyst's high basicity which improved CO_2 adsorption among other factors.

On the other hand, unstable catalysts are also observed. Nascimento *et al.* [407] reported that Pt-based catalysts experienced a reduction in performance within 300 min, due to carbonaceous deposits on sintered metal sites. Also, Co-based catalysts showed continuous reoxidation by steam, water, and CO_2, which also lowers the performance. Kumar *et al.* [404] studied the catalyst stability with different supports on Ni catalysts. Found that MgO supported catalyst underwent re-oxidation, losing active metallic sites despite high resistance to coking due to its basic nature. TiO2-supported catalysts also experienced activity loss due to Ni re-oxidation and sintering, although carbon deposition was minimal due to the negligible acidic nature. CeO_2-ZrO_2 supported catalyst showed high deactivation rates due to Ni sintering, while Al_2O_3 experienced a minor activity loss from surface coking, likely due to the acidic nature Al_2O_3.

The W/F ratio is another important reaction parameter that gives a relation with the amount of catalyst used and the contact time of the feed with the catalyst. A higher catalyst amount increases W/F, while a higher contact time decreases W/F. It is also inversely related to GHSV. Lino *et al.*, and Sun *et al.* [394, 395] observed that both the CH_4 and CO_2 conversions decreased, while the H_2/CO molar ratio increased with higher GHSV. However, beyond a certain GHSV threshold, no significant change was observed. Pandey *et al.* [406] reported negative CO_2 conversion at the lowest GHSV, which increased to a positive value at moderate conditions, but decreased again with a further increase in GHSV. CH_4 conversion consistently decreased with increasing GHSV due to a reduced amount of catalyst and shorter contact time. Walker *et al.* [405] reported no significant change in CH_4 conversion, slightly lower CO_2 conversion, and reduced H_2/CO ratio at higher GHSV.

Conversion rates are related to the contact time between reactants and catalysts, which decreases as GHSV increases [395]. Higher GHSV leads to lower conversion of reactants, due to lower interaction of molecules with catalyst-active sites [394]. Also, GHSV affects the reactor temperature since the heat released in reactions is affected, which again affects CH_4 and CO_2 conversions [395]. While lower catalyst bed temperature favors the WGS reaction and increases the H_2/CO molar ratio [395], higher GHSV and higher temperature favor RWGS reaction and decrease the H_2/CO molar ratio [394]. Higher GHSV may be advantageous for maintaining higher H_2 production by limiting reactions that

consume H_2 [405]. However, at high GHSV, the endothermic SRM and DRM reactions become less important, while methane oxidation becomes predominant, affecting the H_2/CO ratio [406].

Many researchers have proven that there is an optimal threshold for the content of active metal, which varies slightly depending on the catalyst and reaction [395]. Sun *et al.* [395] reported that CH_4 conversion increased as Ni content increased to 10wt%, however, showed no significant change with additional increase. Further, CO_2 conversion decreased significantly, resulting in lower overall catalytic activity compared to 10wt% metal catalyst. Vita *et al.* [396] reported that both CH_4 and CO_2 conversions increased with increased Ni content until the threshold value of 17.6wt%, while Pandey *et al.* [406] reported the threshold value as 5 wt%.

This behavior is primarily due to the increase of active sites and/or surface basicity on the catalyst surface as a direct consequence of the increased amount of active metal in the catalysts [396]. However, increased metal loading weakens the metal-support interaction, reducing the surface basicity compared to catalysts with lower metal content [399]. Despite the increased active metal content, due to the limitations in catalyst synthesis, metal dispersion may deteriorate and clusters may be formed, resulting in decreased catalytic activity [396]. Maximum activity is achieved with catalysts that have the highest metal dispersion and basicity [406]. Additionally, carbon deposition from the Boudouard reaction and methane cracking is more effective on high metal content catalysts, due to a greater number of active metallic sites, as well as the larger crystallite and pore sizes [399].

Promoters in catalytic systems can be classified into two categories; chemical promoters, which enhance the activity of the active species by participating in surface reactions, and textural promoters, which stabilize the catalyst structure and maintain the catalyst activity over time. The chemical promoters and active species used in the experiments within the dataset include Ni, Co, Pt, Rh, Mo, Fe, Ru, and Cu. Some of these metals are used solely as active species in certain studies, others, such as Mo and Co, served also as chemical promoters. Ru and Cu are consistently used only as chemical promoters. The choice of support/carrier and textural promoters is also as critical as an active metal for

catalytic activity. In the dataset, materials such as $MgAl_2O_4$, SBA-15, Al_2O_3, beta-silicate, cordierite, MWCNT, Nb_2O_5, hydrotalcite, TiO_2, SiO_2 are frequently used as carrier materials. MgO, La_2O_3, CeO_2, ZrO_2, Al_2O_3 served dual roles as both support/carrier and textural promoters, while Y_2O_3, KNO_3, Sm_2O_3, ZnO, SrO_2 is used only as textural promoters. Figure 5.3 summarizes the catalyst characteristics used in the dataset.

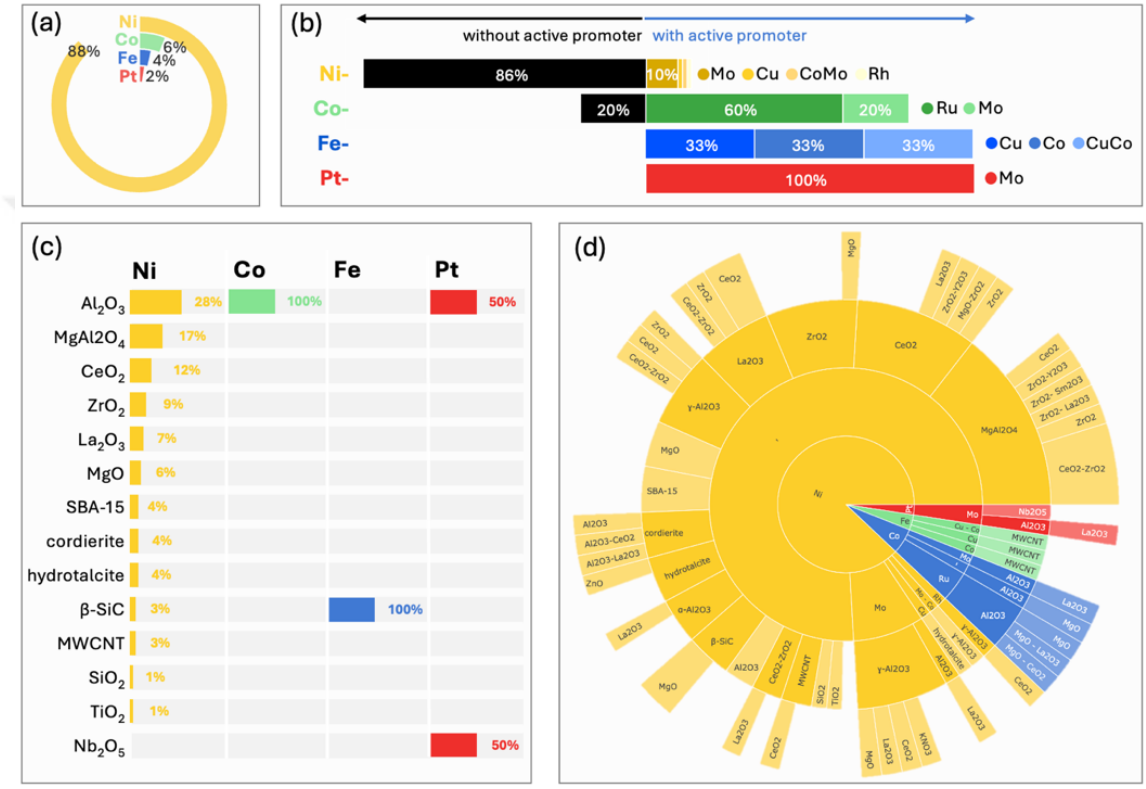

Figure 5.3. Catalyst structures used in the dataset (from center: active metal, chemical promoter, support, textural promoter).

5.2. Machine Learning on Tri-reforming of Methane

In this section, different ML models are explored for their capability of modeling tri-reforming of methane. The selected output variables were; CH_4 conversion, CO_2 conversion, and syngas ratio. The characteristic of the dataset is described in Figure 5.1 and Table 5.1. Two different predictive algorithms; ANN and RF, and classification algorithm; DT is used for model development. Figure 5.4 gives the general framework followed in the study.

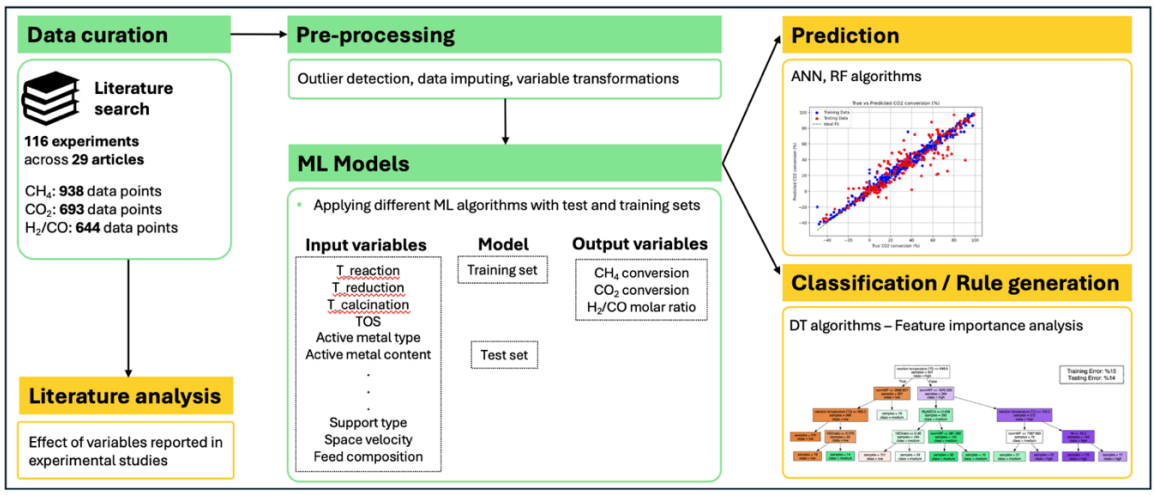

Figure 5.4. General framework of the study.

5.2.1. Computational Details

The ANN models are developed with Python programming language and Keras package. For model optimization and tuning; hidden layer activation function, learning rate, batch size, number of neuron in each layer, and number of layers are investigated. Different activation functions (tanh, sigmoid, relu, softmax, etc.) and different learning rates (0.0001, 0.001, 0.005) are studied. "relu" function with 0.0001 learning rate, and batch size as 80 was found to be the most robust and effective one for all outputs. Data is divided into training and test as 70:30. Optimal model is selected to be not too complex despite of marginal gains.

In RF algorithm, constructed in Python, a 5-fold cross-validation procedure was employed to identify the optimal model parameters. Data is divided into training and test as 70:30. Optimal model is selected to be not too complex despite of marginal gains, which were determined to be *n_estimators*=100 and *min_samples_split*=2.

In DT algorithm, constructed in R, data is divided into three equal classes for each target variable to show low, medium, and high performance; and the path for high CH_4 and CO_2 conversions or desired H_2/CO molar ratio were identified. The CH_4 conversion (%) is divided into three classes as less than 40% (low class), between 40 and 80% (medium class), and more than 80% (high class). The CO_2 conversion (%) is divided as less than 10%,

between 10 and 40%, and more than 40%. Boundaries for CO_2 is lower than CH_4 conversion, as depending on the reactions low CO_2 conversion or even net CO_2 production can take place. The H_2/CO ratio is divided as less than 1.5, between 1.5 and 2.5, and higher than 2.5.

For ML models with missing data, there are three conventional methods that can be used: *(i)* remove the variable, *(ii)* imputation (with some metric), *(iii)* filling the value with some apriori knowledge. For the cases where a lot of missing values are present, removing the variable from the model is performed. For other missing values, before imputing the values with some statistics (e.g., mean) filling the value with some knowledge is performed. Literature is searched to look at other papers published from the same research groups and a detailed search is conducted to fill the necessary information from related publications if it is available. Imputation with median value is performed for *"time on stream"* variable; 60 min. A major apriori knowledge used was to assume the catalyst preparation method as *"wet impregnation"* if the method is only described as *"impregnation"*. Also, it is assumed that if it is not explicitly specified in the articles, the *"H_2% in the reduction medium"* is taken as 100%, and the inert ratio is taken as 0% in the feed composition. Catalysts are tabulated as two parts: an active part, and a support part. Active parts are in weight % whereas in support it is mole %. If the calcination is not specified or performed it is taken as 25 C and 0 hr. W/F is normalized by isolating the inert flow present in the feed flow. Also, flows of reactants (H_2O, O_2, CO_2) are taken as the ratio with respect to CH_4, yielding CH_4 ratio as unity in all samples. For convenience, all alumina types are taken as the same material. Also, active metals are taken as categorical variables, where the total amount of metals are given as numeric variable in another descriptor. TOS is log transformed, as the catalyst activity does not linearly decrease over time.

5.2.2. Results of ANN Analysis

CH_4 conversion model is optimized with 3 hidden layers, 70 neurons in each layer, 180 epochs**.** CO_2 conversion model is tuned with 3 hidden layers, 80 neurons in each layer, 200 epochs. H_2/CO molar ratio model is tuned with 5 hidden layers, 70 neurons in each layer, 200 epochs. Figure 5.5 shows the model R^2 values for different number of neurons in each layer and number of layers for each dataset. R^2 values for training, test, and overall sets

for selected optimum models were 0.936, 0.843, 0.908 for CH_4 conversion, 0.687, 0.623, 0.676 for CO_2 conversion, and 0.858, 0.673, and 0.810 for H_2/CO molar ratio, respectively.

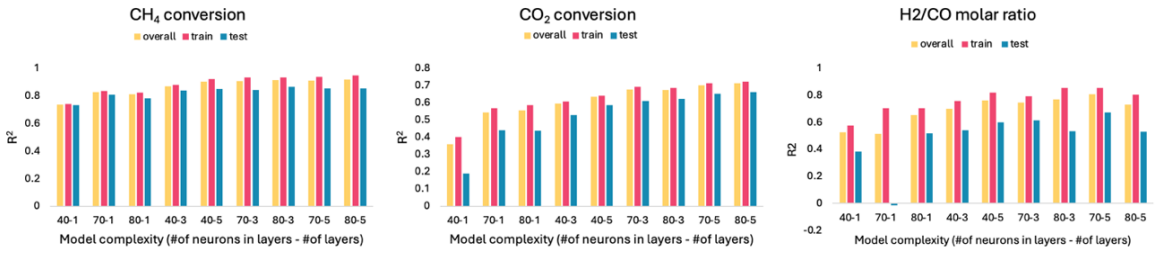

Figure 5.5. ANN Model parameter optimization results of (a) CH_4 conversion (%), (b) CO_2 conversion, (c) H_2/CO molar ratio.

Figure 5.6 shows the experimental versus predicted values for CH_4 conversion (%), CO_2 conversion (%), and H_2/CO molar ratio. Data points from the training set are depicted in blue, while those from the test set are shown in red. To explore the capabilities and shortcomings of the models, highest error values are explored and labeled. In data a&b, no activity is measured in low reaction temperature although model predicts higher conversion. Data c has highest space velocity in dataset, kind of outlier. In data d, TOS has negative and high effect, that model doesn't anticipate, Data e has no active metal: kind of outlier, In data f, effect of increasing TOS results with lower CH_4 conversion; higher than expected. In data g, effect of H_2O, model didn't learn. In data h, catalyst is more stable than expected. And in data i, catalyst is more deactivated than expected.

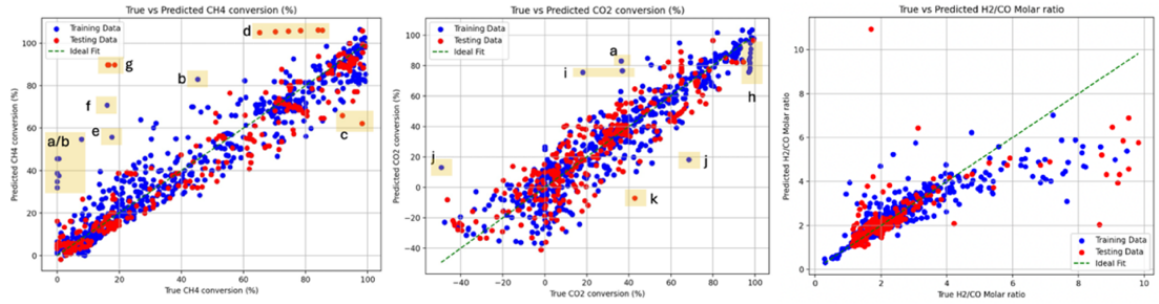

Figure 5.6. Results of ANN models of (a) CH_4 conversion (%), (b) CO_2 conversion, (c) H_2/CO molar ratio.

5.2.3. Results of Random Forest Analysis

Figure 5.7 shows the results of the optimal RF models with experimental versus predicted values for CH_4 conversion (%), CO_2 conversion (%), and H_2/CO molar ratio. Data points from the training set are depicted in blue, while those from the test set are shown in red.

The input significance analysis was also performed for all three target variables. The temperature was the most significant factor as expected; the feed composition was also found to be important (especially for CO_2 conversion) even though its effects, as the effects of other descriptors, seem to be overshadowed by the temperature (at least for CH_4 conversion). Consequently, we decided to discuss the effects of feed composition and other significant descriptors below without referring to the individual importance plots in Figure 5.7.

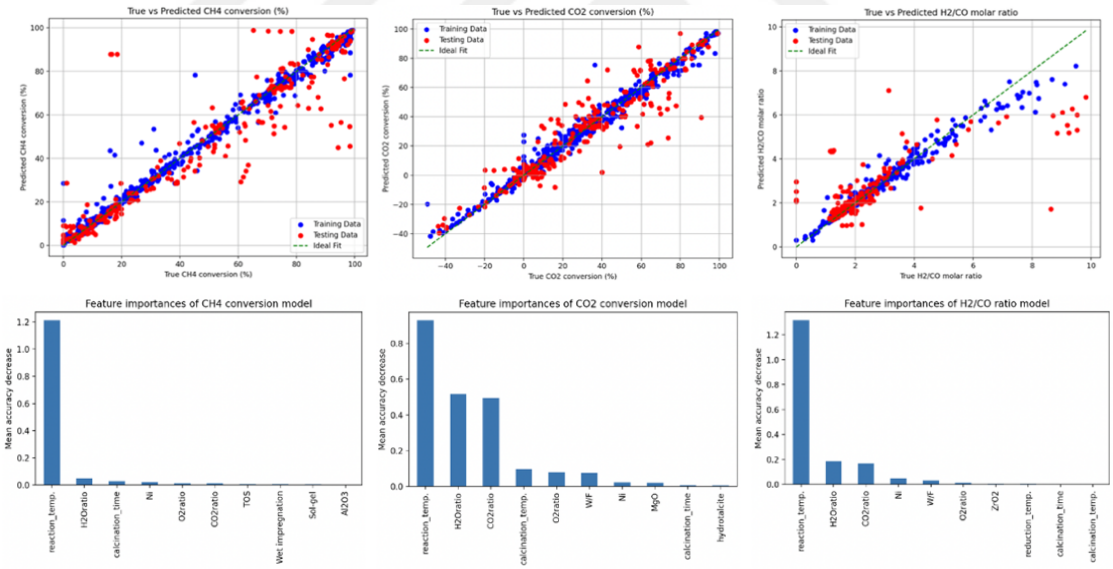

Figure 5.7. RF model results for biogas tri-reforming: (a) CH_4 conversion (%), (b) CO_2 conversion, (c) H_2/CO molar ratio.

Table 5.2 provides statistical metrics for the RF models. By analyzing the R^2 values, it is evident that powerful models were developed for CH_4 and CO_2 conversion. However, the prediction accuracy for the H_2/CO molar ratio was comparatively lower; this is not surprising because H_2/CO ratio is more sensitive than conversion (at least changes with both

CH_4 and CO_2 conversion). The presence of a smaller number of data points may also contribute to this result.

Table 5.2. Results of RF model metrics for biogas tri-reforming process.

		CH_4 conversion (%)	**CO_2 conversion (%)**	**H_2/CO molar ratio**
Train	Mean Squared Error (MSE)	12.87	22.35	0.09
	Root mean squared error (RMSE)	3.59	4.73	0.31
	Mean Absolute Error (MAE)	1.74	2.66	0.16
	R-squared (R^2)	0.988	0.987	0.965
Test	Mean Squared Error (MSE)	150.36	125.46	1.41
	Root mean squared error (RMSE)	12.26	11.20	1.18
	Mean Absolute Error (MAE)	6.30	7.77	0.59
	R-squared (R^2)	0.865	0.908	0.579

5.2.4. Results of Decision Tree Analysis

The discussions presented in the previous section are quite beneficial to understand the effects of individual descriptors on the target variables. However, these descriptors interact with each other and produce the desired outcome with their optimum configuration. In this section, we present a decision tree analysis to develop some heuristic rules that describe the combinations of individual descriptors to be selected for the desired outcome.

The optimized DT models and their corresponding rules are presented in Figure 5.8. DT consists of decision nodes linked by branches, descending from the root node to the terminal leaf nodes. When it is used for classification purposes, tree begins at the root node, located at the top of the diagram, and progresses downward by splitting the data into predefined classes based on the output variable values (e.g., high, medium, low) until reaching the leaf nodes. At each decision point, the branches are separated into either further branches or a terminal leaf to isolate a single class with higher purity. The series of decisions along a branch, leading to a specific class (such as high conversion), can be viewed as a set of rules or heuristics [408].

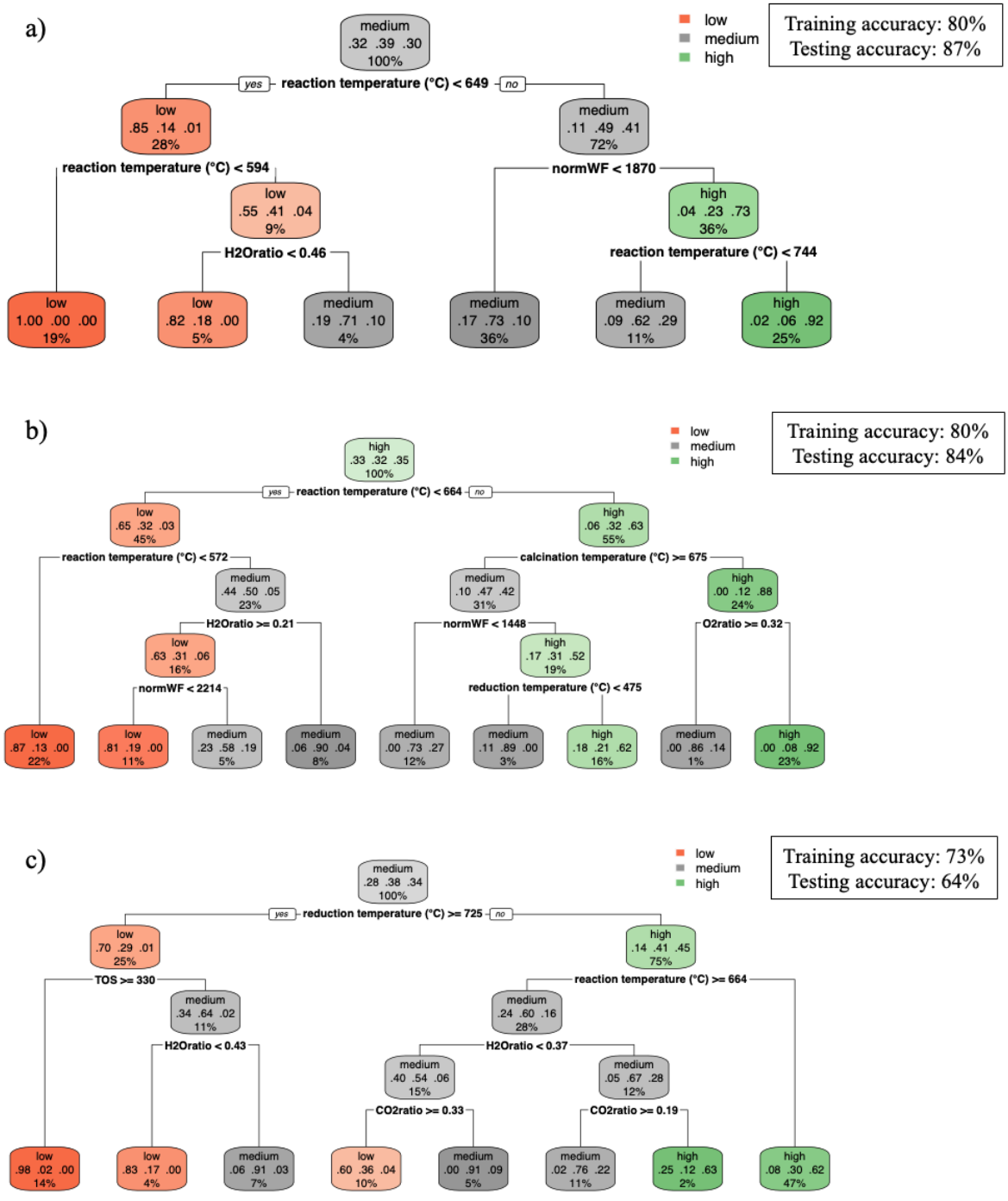

Figure 5.8. Results of DT classification: (a) CH_4 conversion (%), (b) CO_2 conversion, (c) H_2/CO molar ratio.

Notably, a robust rule for high CH_4 conversion class is deduced (the leftmost green branch) , particularly emphasizing the influence of high reaction temperature (>744 °C), and high normalized W/F ratio (>1870 mgcat*min/L). This finding aligns with the literature,

suggesting that elevated reaction temperatures are typically associated with higher conversion rates. This is because carbon formation via the Boudouard reaction from CO and CH_4 cracking becomes less significant at higher temperatures [155]. According to [409], CH_4 cracking, occurs above 557°C, whereas the Boudouard reaction, which generates carbon from CO, occurs below 700°C [155]. Additionally, the flow rate plays a significant role, as reversely related to W/F, which is also stated in the literature, TRM performance tends to improve as feed flow rates decrease [158]. The optimized DT models and their corresponding rules are presented in Figure 5.8.

For CO_2 conversion, high conversion rates are achieved with high reaction temp (>664°C), and by either low calcination temp (<675.0°C), and low O_2 ratio (<0.32) or moderate normalized W/F (>1448 mgcat*min/L) and not low reduction temperature (>475°C). This finding aligns with the literature, as reported by [406], at moderate GHSV the CO_2 conversion achieves maximum. And increase in oxygen amount in the feed decreases CO_2 conversion, favors the methane oxidation route over dry reforming [162, 394-399].

Lower than 1.5 is achieved by high reduction temperature, and TOS higher than 330 min. Between 1.5 and 2.5 is achieved by high reduction temperature, using low reaction temperature, and CO_2 higher than 0.33. A high O_2/CO ratio is achieved by high reduction temperature, high reaction temperature (>650.0°C), and high H_2O ratio (0.26). Feed composition has an important effect on H_2/CO as seen in the DT model. In all rules, not depending on other conditions, increasing H_2O, and decreasing CO_2 in the feed composition increases H_2/CO ratio. This behavior is also observed in the literature, as studies of Vita *et al.* [396, 399, 402] explains the effect of H_2O, studies of Kumar *et al.* [399] and Garcia-Vargas *et al.* [392] backs up the effect of CO_2. suggesting that a higher H_2O/CO_2 ratio favors the SRM and suppresses DRM. Also, the reaction temperature is highly influential to determining the H_2/CO ratio, where lower than 650.0°C temperatures yielded high syngas ratio together with low reduction temperature (<725). This behavior is attributed to promoting WGS reaction and limiting RWGS, which reduces the CO production and increases the H_2/CO ratio [396].

5.3. Transfer Learning on Tri-reforming of Methane

ML models have demonstrated high accuracy in given tasks; however, they require substantial amounts of data. Hence, data scarcity and the high cost of data acquisition especially from laboratory experiments, often limit the predictive power and generalizability of models developed using traditional supervised ML methods [410]. Especially, ML models to be developed for the predictions involving chemical reactions, need extensive labelled datasets to avoid overfitting or underfitting, which can result in low accuracy and transferability [411].

Transfer learning (TL) enables the development of data-driven models for tasks with limited data by sharing information from a related task. This approach is beneficial when (i) data for a particular target task is limited, (ii) there are one or more related tasks with well correlated (though not identical) domains or distributions, and (iii) a substantial amount of data is available for these related tasks. By transferring features, model structure, or parameters from related tasks, the performance of ML models for the target task can be enhanced [410]. To illustrate, Weiss *et al.* [412] used the following analogy: two individuals learning to play the piano; one with no prior musical experience and the other with extensive guitar-playing knowledge. The individual with a music background will learn piano more efficiently by transferring their existing musical knowledge to this new task. Similarly, transfer learning operates on the principle of leveraging knowledge from related tasks to improve learning efficiency on the target task. TL methodology involves pre-training a model on a related task with rich data. In neural network models, it is argued that the initial layers learn fundamental features. To transfer knowledge, the structure and parameters of these initial layers are carried over to the new model, where parameters can either be frozen or fine-tuned. Subsequently, a few additional layers are added, and the model is trained on the limited data of the target task. This approach minimizes the number of parameters that need to be adjusted for the target model [410]. Previously transfer learning is utilized in chemical engineering in applications as; prediction and knowledge extraction of cyanobacterial bloom growth characteristics [413], prediction of temperature-dependent reaction rate constants [414], prediction of molecular properties [410].

In this work, ANN models are developed that use transfer learning methodology to predict CH_4 conversion of TRM reaction, transferring information from SRM and DRM processes, which form some of the reactions in tri-reforming. Also, CO_2 conversion and syngas ratio (H_2/CO) is modeled. For TL studies in biogas tri-reforming process, source datasets are adopted from previous works done by our research group; biogas dry-reforming by Şener *et al.* [176], and steam reforming by Baysal *et al.* [179]. The target (tri-reforming) dataset was described in Section 5.2.

5.3.1. Computational Details

Comprehensive datasets were constructed by reviewing scientific papers sourced from "Web of Science" database. Tri-reforming dataset has a total of 938 data points, originating from 116 experiments across 29 articles as explained in Chapter 5. Steam reforming dataset has a total of 1249 data points, originating from 175 experiments across 41 articles. Dry reforming dataset has a total of 3301 data points, originating from 593 experiments across 101 articles. Detailed explanation of data curation methods and processes for SRM and DRM are explained in Şener *et al.* [176], and Baysal *et al.* [179].

Input variables (descriptors) are chosen as the type of catalyst, catalyst preparation conditions and method, reaction conditions and feed gas compositions, while CH_4, CO_2 conversion (%) and H_2/CO ratio were designated as output (target) variables. Despite the presence of common descriptors (all operational variables; calcination temperature, calcination time, reduction temperature, reduction time, H_2 concentration in the reduction medium, reaction temperature, time on stream (TOS), and CH_4 ratio in the feed stream), there are also differences in the descriptors between the TRM, DRM, and SRM datasets. To make the descriptors across datasets identical, additional descriptors are added with empty values to necessary datasets. For example; some of these descriptors are specific to the tri-reforming reaction, such as no O_2 and H_2O in the reactant medium for the dry reforming, and no CO_2 and O_2 for the steam reforming. Other variables were added to account for different experimental environments, as different base, promoter, and support materials not present in either the source or target datasets. At the end, all datasets are completed to 58 descriptor variables. Some of the variables are also excluded from the dataset as they are not

extensively reported in the articles. Catalyst characteristics (e.g., surface area, pore volume, average pore diameter), some of the catalyst preparation related variables (e.g., dehydration temp. and duration) together with some of the reaction conditions (e.g., reactor type, pressure) are not used in the models. The final variables in the datasets are presented in Table 5.3. Their applicable ranges are presented in the Section 5.1 for TRM and in corresponding articles for SRM and DRM.

Table 5.3. Variables available in dataset and used in models.

Category	**Variables**
active metal	Ni, Co, Pt, Rh, Mo, Fe, Ru, Cu, B
support type	ZnO, La_2O_3, Al_2O_3, $MgAl_2O_4$, MgO, CeO_2, ZrO_2, SrO_2, β-SiC, cordierite, SBA-15, Y_2O_3, KNO_3, Sm_2O_3, Nb_2O_5, MWCNT, TiO_2, hydrotalcite (Mg-Al-C), SiO_2, Gd_2O_3
catalyst preparation conditions	calcination time (h), calcination temperature (°C), H_2% in reduction medium, reduction time (h), reduction temperature (°C)
reaction conditions	time on stream (TOS) (min), W/F, reaction temperature (°C)
feed composition	CH_4%, CO_2% (0 in SRM), H_2O% (0 in DRM), O_2% (0 in DRM, SRM), H_2%, inert %

ANN algorithms are employed for model development in all of the models, including TL studies. First, to establish a benchmark, standard ANN models are developed solely utilizing the TRM dataset for CH_4 conversion, CO_2 conversion, and the syngas ratio. The results of these models serve as benchmarks and compared against to the models incorporating TL techniques. Before introducing TL models, a similar, easier to apply but computationally more demanding models are developed, which are called combined dataset models. These models are developed combining TRM and DRM datasets for CH_4 output variable only, as DRM dataset does not contain other output variables.

Employment of TL is similar to ANN model development, where input variables of the selected dataset are fed to layers of ANN. By estimating the weights for individual nodes in each layer, the algorithm strives to make the best possible prediction for the selected output variable. In each layer, corresponding nodes are combined and activation functions are applied before passing through the following layer. Estimation of weights of each input

variable for each node is done using training and testing sets, and the selection of activation functions are done through trial and error. In TL methodology, an ANN model is developed by using source dataset (DRM or SRM) and saved. Then, another ANN is developed by using the target dataset (TRM), however this time, the computed weights and activation functions of the source model are transferred to the selected layers and locked, meaning that those layers are not tuned with TRM data.

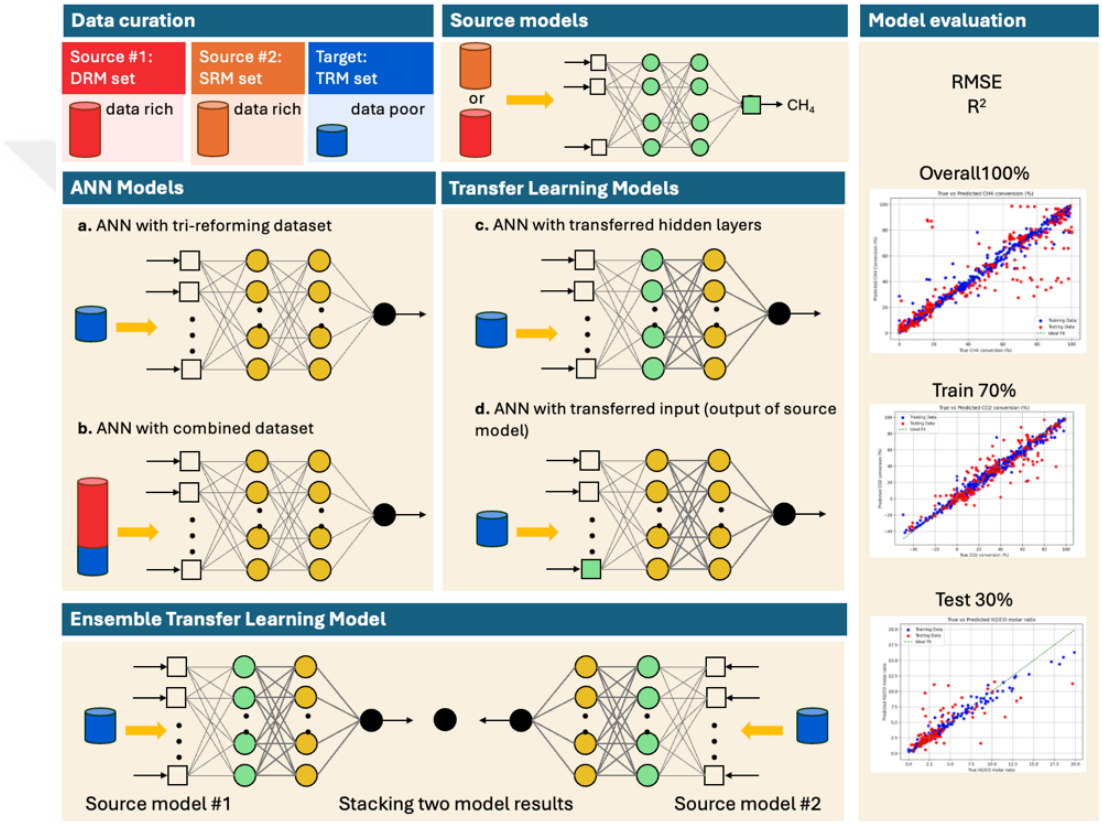

Figure 5.9. Structure of the study for transfer learning approach with ANN.

Different TL configurations are possible, where transferred layers can vary; for instance, transferring only the first layer, the first two layers, or only the last layer from the source model. Layers that are not transferred are independently computed using the target dataset. Another approach to TL involves transferring the knowledge of the source model as an additional input variable to the new model. This approach is employed for CO_2 conversion and H_2/CO ratio models, as the DRM and SRM datasets do not have these output variables. The source ANN model for CH_4 conversion is used to predict CH_4 conversion,

and these predictions are used as an additional input variable for ANN models predicting CO_2 conversion and H_2/CO ratio. Figure 5.9 illustrates the overall scheme followed in ANN models with TL methodology.

ANN models were developed using the Python programming language and the Keras package. For model optimization and tuning, parameters such as activation functions, learning rates, batch sizes, number of neurons per layer, and the number of layers were investigated. Activation functions (tanh, sigmoid, relu, softmax) and learning rates (0.0001, 0.001, 0.005) were tested. The "relu" function with a learning rate of 0.0001 and a batch size of 80 was found to be the most robust and effective across all datasets. Data was split at a ratio of 70:30 for training and testing, except for source model development (80:20). Each layer in all models contained 70 neurons. The base model (using only TRM dataset) was optimized with 3 hidden layers and 180 epochs. Source models (using DRM and SRM dataset) was optimized with 5 hidden layers and 100 epochs. The combined dataset model had 3 hidden layers and 180 epochs. Transfer learning models were also optimized with 180 epochs.

In Section 5.3.2, first, models developed with ANN are compared to each other. Models were developed (i) with solely tri-reforming dataset (benchmark model) (Figure 5.9a), (ii) with combined dataset of tri-reforming and dry-reforming (Figure 5.9b), and (iii) transfer learning models with different transferred layers where dry reforming dataset is used as source model (Figure 5.9c). Then, the effect of number of datapoints used in the model development in the model accuracy is investigated with transfer learning models and compared with benchmark model. Finally, the adoption of two source models to predict the CH_4 conversion is developed and again compared with the benchmark model with different number of datapoints. In Section 5.3.3, models for CO_2 conversion and H_2/CO ratio are developed with ANN models using either solely tri-reforming data, or transfer learning models with dry-reforming dataset. However, as the source model does not contain information about those target variables, the transfer learning methodology was different from CH_4 conversion. Developed model results from the source dataset is used as an additional input variable into the tri-reforming dataset, as shown in Figure 5.9d.

5.3.2. Results of Transfer Learning Models on CH_4 conversion

The study of developing predictive models for CH_4 conversion is divided into three main parts. In the first part, four ANN models were constructed to investigate the effect of transferring different layers. A model developed exclusively from tri-reforming data served as the benchmark model. Subsequently, a model using dry reforming data was developed as the source model. Transfer learning algorithms were then applied by transferring various layers from the source model to predict tri-reforming CH_4 conversion. Additionally, a general model was created by combining both datasets. In the second part of the study, the original tri-reforming dataset was reduced to 75%, 50%, and 25% of its data points. Models were developed to assess the impact of smaller datasets on the performance of both the benchmark model and the TL models. In the last part, steam reforming data was incorporated as a second source model. In conjunction with the dry reforming dataset, the steam reforming dataset was utilized to develop transfer learning models. Two prediction results received from corresponding TL models are used for a new model development for CH_4 conversion prediction to enhance model performance. This double sourced transfer learning models is performed with three different stacking methods; linear regression, gradient boosting regression, and random forest regression.

In the first study, various combinations of hidden layers and transferred layers were analyzed to identify the optimum transfer learning model and compared with benchmark model. Among different combinations tested, optimal transfer learning configurations identified were: (1) 3 new layers with the last layer transferred (3 hidden layers total), (2) the first layer transferred and 3 new layers added (3 hidden layers total), and (3) the first and second layers transferred with 4 new layers added (5 hidden layers total). The models and their R^2 values are presented in Figure 5.10. Data points from the training set are depicted in blue, while those from the test set are shown in red.

Benchmark model developed by using solely the tri-reforming dataset demonstrated the highest predictive power. Transfer learning models varied based on which layer was transferred, with the most successful being the model where the last layer was transferred. Although in ML models the initial layers typically contain the most information and are therefore crucial, transferring these initial layers resulted in decreased overall model

accuracy. It can be concluded that the tri-reforming dataset alone was sufficient to predict the general behavior of CH_4 conversion. Additionally, the model developed using the combined dataset also exhibited good predictive power but requiring higher computational resources.

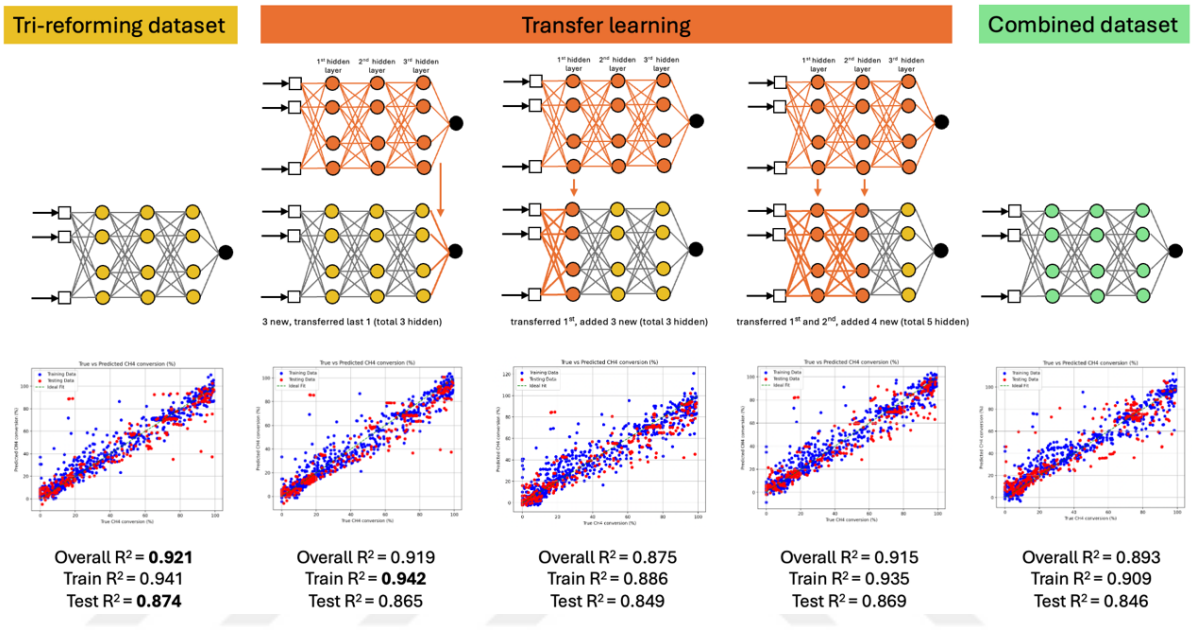

Figure 5.10. Results of CH_4 ANN models.

In the second study, to evaluate the effectiveness of TL, number of data points in the tri-reforming dataset was reduced during model training. Benchmark ANN and TL models were developed using 75%, 50%, and 25% of the tri-reforming dataset. The results indicated that while transfer learning may not be essential for the full dataset (with minor improvements observed when transferring only the last layer) indicating its sufficiency for the tasks, its predictive power becomes significantly higher than only modeling with target model if fewer datapoints are available. TL models are demonstrating significant success even if the data set is reduced. Although transferring the last layer was the most effective TL approach for the complete dataset, this method proved less successful with reduced datasets, as it closely resembled the original model. When data is limited, leveraging knowledge from a source model becomes crucial. As presented in Figure 5.11, the highest accuracy for a reduced dataset (25% of the original size) was achieved by transferring the first two layers from the source model.

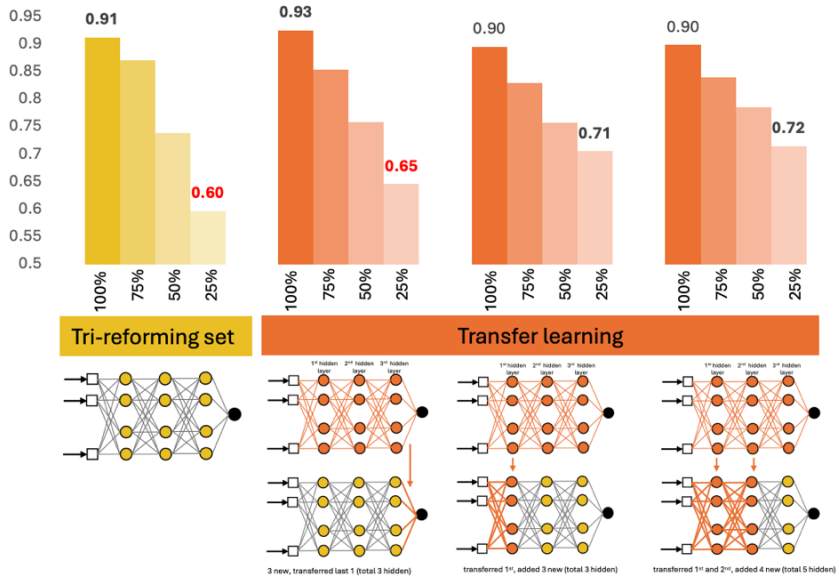

Figure 5.11. Results of CH_4 TL models.

In the third study, as transfer learning models developed with dry-reforming dataset as source model to predict CH_4 conversion from TRM process proved to be useful, especially when the datapoints are limited, use of another source dataset from another major reaction of TRM process is studied: steam reforming of methane (SRM). To develop a predictive transfer learning model for CH_4 conversion using two source models, merging techniques are employed. Merging machine learning models can be accomplished through various methods, with the most common being bagging, boosting, and stacking. In this study, stacking was utilized as a merging technique. Stacking was performed by applying regression models to the final output predictions of the two branches (i.e. ANN models developed using DRM and SRM data are combined). Three different regression models are employed for this purpose; linear regression, gradient boosting regression, and random forest regression. The results are presented in Table 5.4. The base model is built with 3 layers and 70 neurons each. Single and double source TL models are built by transferring first 2 layers from source model and training 3 additional layers.

The results indicate that using two transfer learning model stacked together by using either gradient boosting regression (GBR) or random forest regression (RFR), the predictive power of the model is significantly improved. Even if the 25% data is available for model

development from the tri-reforming dataset, R^2 values were around 0.8 in total dataset. Also, another important result is that the method used for merging (stacking) the two models are also important and achieves different accuracies. It can be observed that the effect of employing a second source model on the overall accuracy of the CH_4 conversion predictions for TRM reaction depends on the stacking method. Although this variability between methods can likely be attributed to many factors, and can only be guessed, most probably reasons are discussed. First, the low number of datapoints in the second source dataset (similar to the target set) may result in favoring this dataset in some of the regression methods. Additionally, the physical conditions of the target model exhibit more DRM characteristics than SRM, as most studies were conducted in low H_2O and high CO_2 environments.

Table 5.4. Results of CH_4 Ensemble transfer learning models.

Data used from tri-reforming set	Benchmark model	Single Source TL Model	Double Source TL Models		
	(only TRM dataset)	(source model: DRM)	(with LR merging)	(with GBR merging)	(with RFR merging)
100%	0.914	0.882	0.882	0.913	**0.934**
75%	0.872	0.856	0.864	0.892	**0.891**
50%	0.740	0.825	0.844	0.848	**0.871**
25%	0.598	0.737	0.729	0.796	**0.804**

5.3.3. Results of Transfer Learning Models on CO_2 conversion and H_2/CO ratio

The source datasets lacked data for CO_2 conversion and H_2/CO ratio, preventing their direct use as source models for transfer learning. Instead, a different transfer learning method was employed for these target variables. For CO_2 conversion and H_2/CO ratio, a base model was built using the tri-reforming dataset. Then TL models are constructed with source models developed using (i) the dry reforming dataset, (ii) the combined dataset, and (iii) the tri-reforming dataset to predict CH_4 conversion. These subsequent models' output then used as an additional input variable for CO_2 conversion and H_2/CO ratio predictions.

CO_2 conversion dataset had fewer datapoints compared to CH_4 conversion dataset, consisting of 693 datapoints. The model was optimized with 3 hidden layers, each containing

80 neurons, a learning rate of 0.0001, the "relu" activation function, 200 epochs, and a batch size of 80. The data was divided into training and testing sets at a ratio of 70:30.

CO_2 conversion model results, presented in Figure 5.12, showed that while the base model was moderately good, transfer learning models performed better. Transfer learning was used to add an input variable (CH_4 conversion) to the prediction models, which improved the accuracy of all TL models compared to the original model. The most successful models were achieved using only dry reforming set as the source model. In contrast, other models, which included the tri-reforming dataset, under-performed, which can be attributed to introduced conflicts.

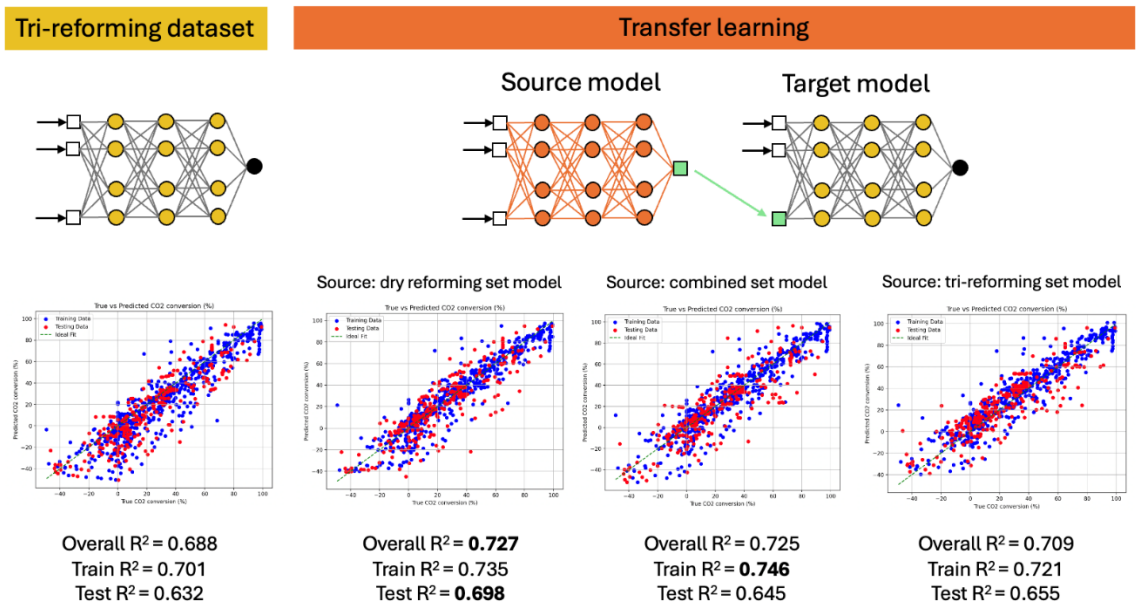

Figure 5.12. Results of CO_2 ANN models.

H_2/CO ratio model showed a similar behavior as CO_2 conversion models as shown in Figure 5.13. This dataset had 644 datapoints. The transferred knowledge from dry reforming dataset significantly improved the predictive power of H_2/CO ratio models, especially for the test set. However, knowledge transfer from datasets including the original tri-reforming dataset is not positively affects the predictions, as the same datapoints already been utilized in base model developments.

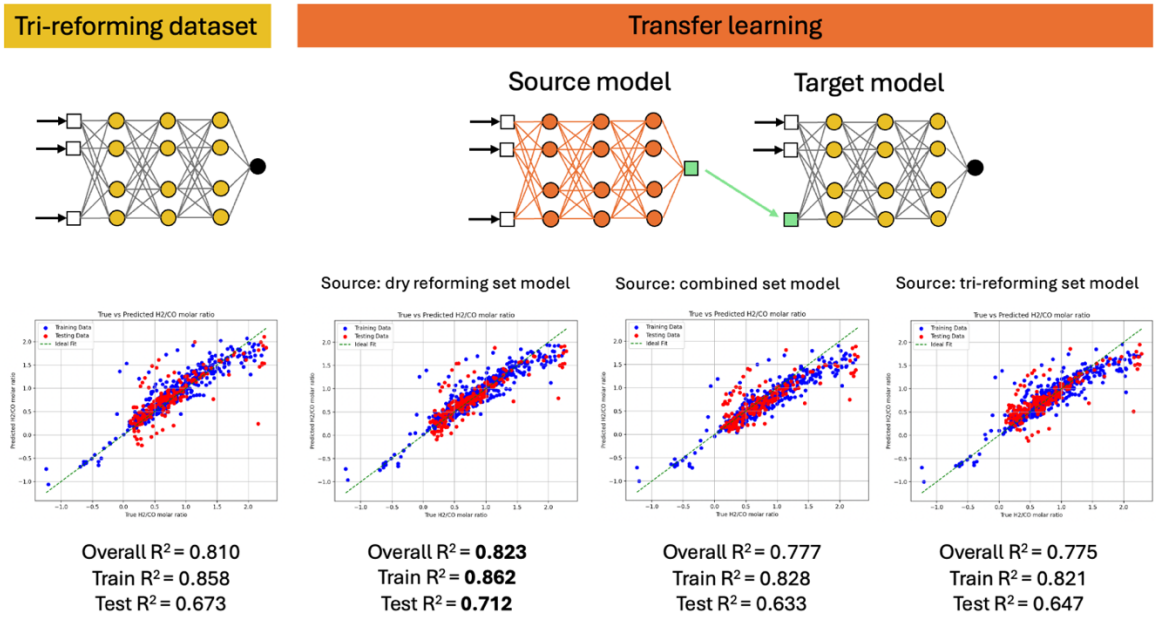

Figure 5.13. Results of H_2/CO ratio ANN models.

6. CONCLUSION AND RECOMMENDATION

To sum up, it is argued that current production strategies make it difficult for biofuels to compete with traditional fuel sources. To achieve the economic viability of biofuels, advanced techniques and concepts have been developed through significant research. The circular economy concept is central to this field, as the production of other value-added chemicals alongside biofuels, cultivation of oleaginous species with mixed cultures, use of wastewater as a cultivation medium, and use of genetically modified strains being discussed for their promising potential in the biodiesel production. Also, use of waste streams and residues, along with mixed feed sources, significantly improves process economics in pyrolysis processes. Additionally, exploring the use of environmentally friendly chemicals and by-products from related bioprocesses is necessary in the processes as transesterification and hydrolysis. These advancements, however, add complexity to an already intricate and nonstandard biological processes. In this study, although machine learning has been applied in this field, many of the described variables have been excluded from dataset. Incorporating these variables is necessary for future research, but the low number of data points spread across a broad range of descriptors poses challenges for implementing machine learning algorithms. Furthermore, unquantified or qualitative variables in experimental studies, such as genetic modifications and waste feed quality, should be quantified for inclusion in machine learning models. These hurdles can be overcome by future advancements in the machine learning, and with recently developed concepts as the use of transfer learning to help learning from other related processes for small datasets, as demonstrated in this dissertation, and hybrid physics-informed models that incorporate simulated data, or physical equations to quantify necessary information for modeling.

6.1. Conclusions

- Biofuel, in general, is a promising alternative to decrease the environmental impacts of fossil fuels; however, they are still not economically competitive with the fossil fuels.

- Microalgae have huge potential for biodiesel production; however, biodiesel produced from microalgae will be likely costlier than the other alternatives. Three courses of action will likely improve the commercial feasibility of the technology: selection of better strains, production of additional value-added chemicals together with lipid and minimization of energy consumption with alternative processes.
- Although lignocellulosic biomass is the most abundant source in nature, converting it to ethanol is not an easy process and involves many sophisticated steps; the pretreatment procedure (i.e. break of cellulosic structure) is one of the most expensive steps with numerous approaches, including physical/physicochemical, acid/alkaline, solvent, and biological treatments.
- In addition to its traditional role as a fuel for some parts of the world, biochar is also critical agent for environmental remediation. Especially functionalized or engineered biochar, which represent the modified biochar for a specific purpose, creates a big opportunity for environmental remediation in the future.
- As in the production of algal biofuels, in the production of biofuel from oleaginous yeasts, one of the most critical step is the selection of suitable strain. For the Yarrowia lipolytica, DT and ARM analysis revealed reliable combinations of variables, such as low moderate C/N molar ratio, limited nitrogen source, and extended growth time for high biomass production. Low C/N molar ratio and low pH were crucial for high lipid content.
- Tri-reforming process for syngas production from biogas is an important research area because biogas production from municipal, agricultural and industrial waste is among the simplest bioprocess, where synthesis gas is an important feedstock to produce synthetic fuels. Analysis of tri-reforming can be summarized in three main areas. First, DT classification proved effective in deriving empirical rules leading to high catalytic performance; these rules may be helpful for the researchers as most TRM works focuses on optimizing catalyst preparation and operational conditions to enhance catalytic performance. It was found that a high reaction temperature combined with a high W/F enhances methane conversion, while a moderate reaction temperature with a low oxygen ratio maximizes CO_2 conversion. Additionally, across all conditions, increasing H_2O and decreasing CO_2 in the feed composition consistently increases the H_2/CO ratio. Second, RF and ANN analysis

with previously performed experimental data is performed and successfully constructed predictive models for CH_4 and CO_2 conversion, and syngas ratio. Lastly, the applicability and potential of transfer learning methods was tested successfully as TRM contains steam and dry reforming reactions which have been studied more frequently (hence have more data on them) for the syngas production; this approach could be also beneficial for the other biofuel processes (for instance cultivation conditions of a new species) that have limited data but similarities with the process (for instance cultivation conditions for a well studies species) with higher amount of available data.

- In general, machine learning was extensively used in biofuel production processes; however, it is a challenge to find sufficient number of data points in these processes because all the species and processes are too complex compared to the material research in which both experimental and computational data are widely available.

6.2. Recommendations

- For algal fuels and oleaginous yeasts, the most critical step is to screen the promising strains and select the most suitable one for the conditions used (*e.g.*, the best algal strain is highly depending on the wastewater to be used as nutrition source). ML can make significant contribution for this even though the construction of a sufficient data set is a challenge. The future studies can be focus on this.
- Transfer learning algorithms may be also useful in various biofuel processes because some steps for the processing of different feedstock or production of different products are similar; the data from more frequently studied fields may be used in new alternatives with lack of sufficient data.
- New ML approaches such as explainable machine learning, physics-informed machine learning or algorithms based on natural language processing could be utilized.

REFERENCES

1. Rodionova, M.V., et al., "Biofuel Production: Challenges and Opportunities", *International Journal of Hydrogen Energy*, Vol. 42, No. 12, pp. 8450-8461, 2017.

2. Adams, P., et al., *Greenhouse Gases Balances of Bioenergy Systems*, Elsevier, London, 2018.

3. Zhang, D., et al., "Analysis of the Cyanobacterial Hydrogen Photoproduction Process via Model Identification and Process Simulation", *Chemical Engineering Science*, Vol. 128, pp. 130-146, 2015.

4. Abdel-Raouf, N., A.A. Al-Homaidan, and I.B. Ibraheem, "Microalgae and Wastewater Treatment", *Saudi Journal of Biological Sciences*, Vol. 19, No. 3, pp. 257-275, 2012.

5. Otálora, P., et al., "Microalgae Classification Based on Machine Learning Techniques", *Algal Research*, Vol. 55, 2021.

6. Brennan, L. and P. Owende, "Biofuels from Microalgae - A Review of Technologies for Production, Processing, and Extractions of Biofuels and Co-Products", *Renewable and Sustainable Energy Reviews*, Vol. 14, No. 2, pp. 557-577, 2010.

7. Gouveia, L., et al., "Microalgae – Source of Natural Bioactive Molecules as Functional Ingredients", *Food Science & Technology Bulletin: Functional Foods*, Vol. 7, No. 2, pp. 21-37, 2010.

8. Bradford, E., et al., "Dynamic Modeling and Optimization of Sustainable Algal Production with Uncertainty Using Multivariate Gaussian Processes", *Computers & Chemical Engineering*, Vol. 118, pp. 143-158, 2018.

9. Pulz, O. and W. Gross, "Valuable Products from Biotechnology of Microalgae", *Applied Microbiology and Biotechnology*, Vol. 65, No. 6, pp. 635-48, 2004.

10. Franco, B.M., et al., "Monoalgal and Mixed Algal Cultures Discrimination by Using an Artificial Neural Network", *Algal Research*, Vol. 38**,** 2019.

11. Levasseur, W., V. Pozzobon, and P. Perré, "Green Microalgae in Intermittent Light: A Meta-Analysis Assisted by Machine Learning", *Journal of Applied Phycology*, Vol. 34, No. 1, pp. 135-158, 2021.

12. Zhang, D., et al., "Hybrid Physics-Based and Data-Driven Modeling for Bioprocess Online Simulation and Optimization", *Biotechnology and Bioengineering*, Vol. 116, No. 11, pp. 2919-2930, 2019.

13. Singh, S.P. and P. Singh, "Effect of CO2 Concentration on Algal Growth: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 38, pp. 172-179, 2014.

14. Kumar, M., et al., "Algae as Potential Feedstock for the Production of Biofuels and Value-Added Products: Opportunities and Challenges", *Science of the Total Environment*, Vol. 716, pp. 137116, 2020.

15. Rani, L., et al., *Artificial Intelligence for Renewable Energy Systems*, Woodhead Publishing, Cambridge, 2022.

16. Suganya, T., et al., "Macroalgae and Microalgae as a Potential Source for Commercial Applications along with Biofuels Production: A Biorefinery Approach", *Renewable and Sustainable Energy Reviews*, Vol. 55, pp. 909-941, 2016.

17. Trivedi, J., et al., "Algae Based Biorefinery - How to Make Sense?", *Renewable and Sustainable Energy Reviews*, Vol. 47, pp. 295-307, 2015.

18. Lee, O.K. and E.Y. Lee, "Sustainable Production of Bioethanol from Renewable Brown Algae Biomass", *Biomass and Bioenergy*, Vol. 92, pp. 70-75, 2016.

19. Adeniyi, O.M., U. Azimov, and A. Burluka, "Algae Biofuel: Current Status and Future Applications"*, Renewable and Sustainable Energy Reviews*, Vol. 90, pp. 316-335, 2018.

20. Joon Sung, Y., et al., "Microalgae-Derived Hydrogen Production Towards Low Carbon Emissions via Large-Scale Outdoor Systems", *Bioresource Technology*, pp. 128134, 2022.

21. Coşgun, A., M.E. Günay, and R. Yıldırım, "Exploring the Critical Factors of Algal Biomass and Lipid Production for Renewable Fuel Production by Machine Learning", *Renewable Energy*, Vol. 163, pp. 1299-1317, 2021.

22. Han, F., et al., "Optimization and Lipid Production Enhancement of Microalgae Culture by Efficiently Changing the Conditions along with the Growth-State", *Energy Conversion and Management*, Vol. 90, pp. 315-322, 2015.

23. Pandey, A., et al., *Biofuels from Algae*, Elsevier, Burlington, 2014.

24. Pushpakumari Kudahettige, N., J. Pickova, and F.G. Gentili, "Stressing Algae for Biofuel Production: Biomass and Biochemical Composition of Scenedesmus dimorphus and Selenastrum minutum Grown in Municipal Untreated Wastewater", *Frontiers in Energy Research*, Vol. 6, 2018.

25. Coşgun, A., M.E. Günay, and R. Yıldırım, "Machine Learning for Algal Biofuels: A Critical Review and Perspective for the Future", *Green Chemistry*, Vol. 25, No. 9, pp. 3354-3373, 2023.

26. Griffiths, M.J., R.P. van Hille, and S.T.L. Harrison, "Lipid Productivity, Settling Potential and Fatty Acid Profile of 11 Microalgal Species Grown Under Nitrogen

Replete and Limited Conditions", *Journal of Applied Phycology*, Vol. 24, No. 5, pp. 989-1001, 2011.

27. Sanyano, N., P. Chetpattananondh, and S. Chongkhong, "Coagulation-Flocculation of Marine Chlorella Sp. for Biodiesel Production", *Bioresource Technology*, Vol. 147, pp. 471-476, 2013.

28. Show, K.-Y. and D.-J. Lee, "Algal Biomass Harvesting", *Biofuels from Algae*, pp. 85-110, 2014.

29. Gerardo, M.L., et al., "Harvesting of Microalgae within a Biorefinery Approach: A Review of the Developments and Case Studies from Pilot-Plants", *Algal Research*, Vol. 11, pp. 248-262, 2015.

30. Mata, T.M., A.A. Martins, and N.S. Caetano, "Microalgae for Biodiesel Production and other Applications: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 14, No. 1, pp. 217-232, 2010.

31. Barros, A.I., et al., "Harvesting Techniques Applied to Microalgae: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 41, pp. 1489-1500, 2015.

32. Balasubramanian, R.K., T.T. Yen Doan, and J.P. Obbard, "Factors Affecting Cellular Lipid Extraction from Marine Microalgae", *Chemical Engineering Journal*, Vol. 215-216, pp. 929-936, 2013.

33. Guldhe, A., et al., "Efficacy of Drying and Cell Disruption Techniques on Lipid Recovery from Microalgae for Biodiesel Production", *Fuel*, Vol. 128, pp. 46-52, 2014.

34. Widjaja, A., C.-C. Chien, and Y.-H. Ju, "Study of Increasing Lipid Production from Fresh Water Microalgae Chlorella Vulgaris", *Journal of the Taiwan Institute of Chemical Engineers*, Vol. 40, No. 1, pp. 13-20, 2009.

35. Ferrell, J. and V. Sarisky-Reed, National Algal Biofuels Technology Roadmap, United States, 2010.

36. Taher, H., et al., "Effective Extraction of Microalgae Lipids from Wet Biomass for Biodiesel Production", *Biomass and Bioenergy*, Vol. 66, pp. 159-167, 2014.

37. Yoo, G., et al., "An Effective, Cost-Efficient Extraction Method of Biomass from Wet Microalgae with a Functional Polymeric Membrane", *Green Chemistry*, Vol. 16, No. 1, pp. 312-319, 2014.

38. Mubarak, M., A. Shaija, and T.V. Suchithra, "A Review on the Extraction of Lipid from Microalgae for Biodiesel Production", *Algal Research*, Vol. 7, pp. 117-123, 2015.

39. Wan Mahmood, W.M.A., C. Theodoropoulos, and M. Gonzalez-Miquel, "Enhanced Microalgal Lipid Extraction Using Bio-Based Solvents for Sustainable Biofuel Production", *Green Chemistry*, Vol. 19, No. 23, pp. 5723-5733, 2017.

40. Halim, R., et al., "Oil Extraction from Microalgae for Biodiesel Production", *Bioresource Technology*, Vol. 102, No. 1, pp. 178-85, 2011.

41. Li, Y., et al., "A Comparative Study: the Impact of Different Lipid Extraction Methods on Current Microalgal Lipid Research", *Microbial Cell Factory*, Vol. 13, No. 14, pp. 1-9, 2014.

42. Ramluckan, K., K.G. Moodley, and F. Bux, "An Evaluation of the Efficacy of Using Selected Solvents for the Extraction of Lipids from Algal Biomass by the Soxhlet Extraction Method", *Fuel*, Vol. 116, pp. 103-108, 2014.

43. Soh, L. and J. Zimmerman, "Biodiesel Production: the Potential of Algal Lipids Extracted with Supercritical Carbon Dioxide", *Green Chemistry*, Vol. 13, No. 6, 2011.

44. Lu, W., Z. Wang, and Z. Yuan, "Characteristics of lipid extraction from Chlorella sp. cultivated in outdoor raceway ponds with mixture of ethyl acetate and ethanol for biodiesel production", *Bioresource Technology*, Vol. 191, pp. 433-437, 2015.

45. Macías-Sánchez, M.D., et al., "Supercritical Fluid Extraction of Carotenoids and Chlorophyll-a from Synechococcus Sp.", *The Journal of Supercritical Fluids*, Vol. 39, No. 3, pp. 323-329, 2007.

46. Wu, J., et al., "Enhanced Extraction of Lipids from Microalgae with Eco-Friendly Mixture of Methanol and Ethyl Acetate for Biodiesel Production", *Journal of the Taiwan Institute of Chemical Engineers*, Vol. 71, pp. 323-329, 2017.

47. Mubarak, M., A. Shaija, and T.V. Suchithra, "Ultrasonication: An Effective Pre-Treatment Method for Extracting Lipid from Salvinia Molesta for Biodiesel Production", *Resource-Efficient Technologies*, Vol. 2, No. 3, pp. 126-132, 2016.

48. Doshi, A., et al., "Economic and Policy Issues in the Production of Algae-Based Biofuels: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 64, pp. 329-337, 2016.

49. Hoffman, J., et al., "Techno-Economic Assessment of Open Microalgae Production Systems", *Algal Research*, Vol. 23, pp. 51-57, 2017.

50. Klein, B. and R. Davis, Algal Biomass Production via Open Pond Algae Farm Cultivation: 2021 State of Technology and Future Research, 2022.

51. Xin, C., et al., "Waste-To-Biofuel Integrated System and its Comprehensive Techno-Economic Assessment in Wastewater Treatment Plants", *Bioresource Technology*, Vol. 250, pp. 523-531, 2018.

52. Hossain, N., et al., "Techno-Economics and Sensitivity Analysis of Microalgae as Commercial Feedstock for Bioethanol Production", *Environmental Progress & Sustainable Energy*, Vol. 38, No. 5, 2019.

53. Subramanian, S. and R.T. Sayre, "The Right Stuff; Realizing the Potential for Enhanced Biomass Production in Microalgae", *Frontiers in Energy Research*, Vol. 10, 2022.

54. Hannon, M., et al., "Biofuels from Algae: Challenges and Potential", *Biofuels*, Vol. 1, No. 5, pp. 763-784, 2010.

55. Chia, S.R., et al., "Algae as Potential Feedstock for Various Bioenergy Production", *Chemosphere*, Vol. 287, No. 1, pp. 131944, 2022.

56. Fayyaz, M., et al., "Genetic Engineering of Microalgae for Enhanced Biorefinery Capabilities", *Biotechnol Advances*, Vol. 43, pp. 107554, 2020.

57. Kim, S.-K., *Handbook of Marine Microalgae*, Academic Press, London, 2015.

58. Mofijur, M., et al., "Selection of Microalgae Strains for Sustainable Production of Aviation Biofuel", *Bioresource Technology*, Vol. 345, pp. 126408, 2022.

59. Nagarajan, D., et al., "Resource Recovery from Wastewaters Using Microalgae-Based Approaches: A Circular Bioeconomy Perspective", *Bioresource Technology*, Vol. 302, pp. 122817, 2020.

60. Brandt, A., et al., "Deconstruction of Lignocellulosic Biomass with Ionic Liquids", *Green Chemistry*, Vol. 15, No. 3, 2013.

61. Kumar, R., et al., "Recent Updates on Lignocellulosic Biomass Derived Ethanol - A Review", *Biofuel Research Journal*, Vol. 3, No. 1, pp. 347-356, 2016.

62. Liu, C.G., et al., "Cellulosic Ethanol Production: Progress, Challenges and Strategies for Solutions", *Biotechnology Advances*, Vol. 37, No. 3, pp. 491-504, 2019.

63. Coşgun, A., M.E. Günay, and R. Yıldırım, "A Critical Review of Machine Learning for Lignocellulosic Ethanol Production via Fermentation Route", *Biofuel Research Journal*, Vol. 10, No. 2, pp. 1859-1875, 2023.

64. Aui, A., Y. Wang, and M. Mba-Wright, "Evaluating the Economic Feasibility of Cellulosic Ethanol: A Meta-Analysis of Techno-Economic Analysis Studies", *Renewable and Sustainable Energy Reviews*, Vol. 145, 2021.

65. Griffin, D.W. and M.A. Schultz, "Fuel and Chemical Products from Biomass Syngas: A Comparison of Gas Fermentation to Thermochemical Conversion Routes", *Environmental Progress & Sustainable Energy*, Vol. 31, No. 2, pp. 219-224, 2012.

66. Li, W., et al., "Comparative Techno-Economic, Uncertainty and Life Cycle Analysis of Lignocellulosic Biomass Solvent Liquefaction and Sugar Fermentation to Ethanol", *ACS Sustainable Chemistry & Engineering*, Vol. 6, No. 12, pp. 16515-16524, 2018.

67. Zoghlami, A. and G. Paes, "Lignocellulosic Biomass: Understanding Recalcitrance and Predicting Hydrolysis", *Frontiers in Chemistry*, Vol. 7, No. 874, 2019.

68. Kumar, S., et al., "Cellulose Pretreatment in Subcritical Water: Effect of Temperature on Molecular Structure and Enzymatic Reactivity", *Bioresource Technology*, Vol. 101, No. 4, pp. 1337-1347, 2010.

69. Cheah, W.Y., et al., "Pretreatment Methods for Lignocellulosic Biofuels Production: Current Advances, Challenges and Future Prospects", *Biofuel Research Journal*, Vol. 7, No. 1, pp. 1115-1127, 2020.

70. Charte, F., et al., "Comparative Analysis of Data Mining and Response Surface Methodology Predictive Models for Enzymatic Hydrolysis of Pretreated Olive Tree Biomass", *Computers & Chemical Engineering*, Vol. 101, pp. 23-30, 2017.

71. Yang, B. and C.E. Wyman, "Pretreatment: The Key to Unlocking Low-Cost Cellulosic Ethanol", *Biofuels, Bioproducts and Biorefining*, Vol. 2, No. 1, pp. 26-40, 2008.

72. Aguilar-Reynosa, A., et al., "Microwave Heating Processing as Alternative of Pretreatment in Second-Generation Biorefinery: An Overview", *Energy Conversion and Management*, Vol. 136, pp. 50-65, 2017.

73. Yan, L., et al., "Hot Water Pretreatment of Lignocellulosic Biomass: An Effective and Environmentally Friendly Approach to Enhance Biofuel Production", *Chemical Engineering & Technology*, Vol. 39, No. 10, pp. 1759-1770, 2016.

74. Liu, C.G., et al., "Assessment and Regression Analysis on Instant Catapult Steam Explosion Pretreatment of Corn Stover", *Bioresource Technology*, Vol. 166, pp. 368-372, 2014.

75. Loow, Y.-L., et al., "Typical Conversion of Lignocellulosic Biomass into Reducing Sugars Using Dilute Acid Hydrolysis and Alkaline Pretreatment", *Cellulose*, Vol. 23, No. 3, pp. 1491-1520, 2016.

76. Xu, Z. and F. Huang, "Pretreatment Methods for Bioethanol Production", *Applied Biochemistry and Biotechnology*, Vol. 174, No. 1, pp. 43-62, 2014.

77. Zheng, Y., Z. Pan, and R. Zhang, "Overview of Biomass Pretreatment for Cellulosic Ethanol Production", *International Journal of Agricultural and Biological Engineering*, Vol. 2, No. 3, pp. 51, 2009.

78. Kim, J.S., Y.Y. Lee, and T.H. Kim, "A Review on Alkaline Pretreatment Technology for Bioconversion of Lignocellulosic Biomass", *Bioresource Technology*, Vol. 199, pp. 42-48, 2016.

79. Galbe, M. and O. Wallberg, "Pretreatment for Biorefineries: A Review of Common Methods for Efficient Utilisation of Lignocellulosic Materials", *Biotechnology for Biofuels and Bioproducts*, Vol. 12, pp. 294, 2019.

80. Sindhu, R., P. Binod, and A. Pandey, "Biological Pretreatment of Lignocellulosic Biomass - An Overview", *Bioresource Technology*, Vol. 199, pp. 76-82, 2016.

81. Ravindran, R. and A.K. Jaiswal, "A Comprehensive Review on Pre-Treatment Strategy for Lignocellulosic Food Industry Waste: Challenges and Opportunities", *Bioresource Technology*, Vol. 199, pp. 92-102, 2016.

82. Vollmer, N.I., et al., "Model Development for the Optimization of Operational Conditions of the Pretreatment of Wheat Straw", *Chemical Engineering Journal*, Vol. 430, 2022.

83. Lugani, Y., et al., "Recent Advances in Bioethanol Production from Lignocelluloses: A Comprehensive Review with a Focus on Enzyme Engineering and Designer Biocatalysts", *Biofuel Research Journal*, Vol. 7, No. 4, pp. 1267-1295, 2020.

84. Vani, S., R.K. Sukumaran, and S. Savithri, "Prediction of Sugar Yields During Hydrolysis of Lignocellulosic Biomass Using Artificial Neural Network Modeling", *Bioresource Technology*, Vol. 188, pp. 128-35, 2015.

85. Agrawal, R., et al., "Current Understanding of the Inhibition Factors and Their Mechanism of Action for the Lignocellulosic Biomass Hydrolysis", *Bioresource Technology*, Vol. 332, pp. 125042 ,2021.

86. Zhao, F., Y. Yi, and X. Lü, *Advances in 2nd Generation of Bioethanol Production*, Woodhead Publishing, Duxford, 2021.

87. Scheller, H.V. and P. Ulvskov, "Hemicelluloses", *Annual Review of Plant Biology*, Vol. 61, pp. 263-289, 2010.

88. Kucharska, K., et al., "Key Issues in Modeling and Optimization of Lignocellulosic Biomass Fermentative Conversion to Gaseous Biofuels", *Renewable Energy*, Vol. 129, pp. 384-408, 2018.

89. Kim, J.H., D.E. Block, and D.A. Mills, "Simultaneous Consumption of Pentose and Hexose Sugars: An Optimal Microbial Phenotype for Efficient Fermentation of Lignocellulosic Biomass", *Applied Microbiology and Biotechnology*, Vol. 88, No. 5, pp. 1077-1085, 2010.

90. Jahanbakhshi, A. and R. Salehi, "Processing Watermelon Waste Using Saccharomyces Cerevisiae Yeast and the Fermentation Method for Bioethanol Production", *Journal of Food Process Engineering*, Vol. 42, No. 7, 2019.

91. Qiao, J., et al., "Integrated Biorefinery Approaches for the Industrialization of Cellulosic Ethanol Fuel", *Bioresource Technology*, Vol. 360, pp. 127516, 2022.

92. Sidana, A. and S.K. Yadav, "Recent Developments in Lignocellulosic Biomass Pretreatment with a Focus on Eco-Friendly, Non-Conventional Methods", *Journal of Cleaner Production*, Vol. 335, 2022.

93. Culaba, A.B., et al., "Smart Sustainable Biorefineries for Lignocellulosic Biomass", *Bioresource Technology*, Vol. 344, pp. 126215, 2022.

94. Smuga-Kogut, M., et al., "Use of Machine Learning Methods for Predicting Amount of Bioethanol Obtained from Lignocellulosic Biomass with the Use of Ionic Liquids for Pretreatment", *Energies*, Vol. 14, No. 1, 2021.

95. Sousa Jr, R., et al., "Recent Trends in the Modeling of Cellulose Hydrolysis", *Brazilian Journal of Chemical Engineering*, Vol. 28, No. 4, pp. 545-564, 2011.

96. Wang, Z., et al., "The Role of Machine Learning to Boost the Bioenergy and Biofuels Conversion", *Bioresource Technology*, Vol. 343, pp. 126099, 2022.

97. Peng, W. and O. Karimi Sadaghiani, "A Review on the Applications of Machine Learning and Deep Learning in Agriculture Section for the Production of Crop Biomass Raw Materials", *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, Vol. 45, No. 3, pp. 9178-9201, 2023.

98. Niaze, A.A., et al., "Model Construction and Optimization for Raising the Concentration of Industrial Bioethanol Production by Using a Data-Driven ANN Model", *Renewable Energy*, Vol. 216, 2023.

99. İşçen, A., et al., "Exploring the Critical Factors of Biomass Pyrolysis for Sustainable Fuel Production by Machine Learning", *Sustainability*, Vol. 15, No. 20, 2023.

100. Zhu, X., Y. Li, and X. Wang, "Machine Learning Prediction of Biochar Yield and Carbon Contents in Biochar Based on Biomass Characteristics and Pyrolysis Conditions", *Bioresource Technology*, Vol. 288, pp. 121527, 2019.

101. Rose, J., et al., "The Forgotten Coal: Charcoal Demand in Sub-Saharan Africa", *World Development Perspectives*, Vol. 25, 2022.

102. Anand, A., et al., "Biochar Production from Crop Residues, Its Characterization and Utilization for Electricity Generation in India", *Journal of Cleaner Production*, Vol. 368, 2022.

103. Tripathi, M., J.N. Sahu, and P. Ganesan, "Effect of Process Parameters on Production of Biochar from Biomass Waste Through Pyrolysis: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 55, pp. 467-481, 2016.

104. Zhang, X. and R.C. Brown, *Thermochemical Processing of Biomass*, Wiley, West Sussex, 2019.

105. Meyer, S., B. Glaser, and P. Quicker, "Technical, Economical, and Climate-Related Aspects of Biochar Production Technologies: A Literature Review", *Environmental Science & Technology*, Vol. 45, No. 22, pp. 9473-9483, 2011.

106. Pang, S., "Advances in Thermochemical Conversion of Woody Biomass to Energy, Fuels and Chemicals", *Biotechnology Advances*, Vol. 37, No. 4, pp. 589-597, 2019.

107. Jeyasubramanian, K., et al., "A Complete Review on Biochar: Production, Property, Multifaceted Applications, Interaction Mechanism and Computational Approach", *Fuel*, Vol. 292, 2021.

108. Peixoto Basso, T., T.O. Basso, L.C. Basso, *Biotechnological Applications of Biomass*, IntechOpen, 2021.

109. Vuppaladadiyam, A.K., et al., "A Critical Review on Biomass Pyrolysis: Reaction Mechanisms, Process Modeling and Potential Challenges", *Journal of the Energy Institute*, Vol. 108, 2023.

110. Venderbosch, R.H., "Fast Pyrolysis", *Thermochemical Processing of Biomass*, pp. 175-206, 2019.

111. Guedes, R.E., A.S. Luna, and A.R. Torres, "Operating Parameters for Bio-Oil Production in Biomass Pyrolysis: A Review", *Journal of Analytical and Applied Pyrolysis*, Vol. 129, pp. 134-149, 2018.

112. Venderbosch, R.H. and W. Prins, "Fast Pyrolysis Technology Development", *Biofuels, Bioproducts and Biorefining*, Vol. 4, No. 2, pp. 178-208, 2010.

113. Woolf, D., et al., "Biofuels from Pyrolysis in Perspective: Trade-Offs Between Energy Yields and Soil-Carbon Additions", *Environmental Science & Technology*, Vol. 48, No. 11, pp. 6492-6499, 2014.

114. Alabdrabalnabi, A., R. Gautam, and S. Mani Sarathy, "Machine Learning to Predict Biochar and Bio-Oil Yields from Co-Pyrolysis of Biomass and Plastics", *Fuel*, Vol. 328, 2022.

115. Shahbeik, H., et al., "Characterizing Sludge Pyrolysis by Machine Learning: Towards Sustainable Bioenergy Production from Wastes", *Renewable Energy*, Vol. 199, pp. 1078-1092, 2022.

116. Uzoma, K.C., et al., "Effect of Cow Manure Biochar on Maize Productivity Under Sandy Soil Condition", *Soil Use and Management*, Vol. 27, No. 2, pp. 205-212, 2011.

117. Huang, S.W., et al., "Bio-Reduction and Synchronous Removal of Hexavalent Chromium from Aqueous Solutions Using Novel Microbial Cell/Algal-Derived Biochar Particles: Turning an Environmental Problem into an Opportunity", *Bioresource Technology*, Vol. 309, pp. 123304, 2020.

118. Koltowski, M., et al., "Effect of Biochar Activation by Different Methods on Toxicity of Soil Contaminated by Industrial Activity", *Ecotoxicology and Environmental Safety*, Vol. 136, pp. 119-125, 2017.

119. Ren, X., et al., "Challenges and Opportunities in Microwave-Assisted Catalytic Pyrolysis of Biomass: A Review", *Applied Energy*, Vol. 315, 2022.

120. Guo, H., et al., "Production of High-Quality Pyrolysis Product by Microwave–Assisted Catalytic Pyrolysis of Wood Waste and Application of Biochar", *Arabian Journal of Chemistry*, Vol. 16, No. 8, 2023.

121. Borges, F.C., et al., "Fast Microwave Assisted Pyrolysis of Biomass Using Microwave Absorbent", *Bioresource Technology*, Vol. 156, p. 267-274, 2014.

122. Mohamed, B.A., et al., "Engineered Biochar from Microwave-Assisted Catalytic Pyrolysis of Switchgrass for Increasing Water-Holding Capacity and Fertility of Sandy Soil", *Science of the Total Environment*, Vol. 566-567, pp. 387-397, 2016.

123. Akhtar, J. and N. Saidina Amin, "A Review on Operating Parameters for Optimum Liquid Oil Yield in Biomass Pyrolysis", *Renewable and Sustainable Energy Reviews*, Vol. 16, No. 7, pp. 5101-5109, 2012.

124. Hu, X. and M. Gholizadeh, "Biomass Pyrolysis: A Review of the Process Development and Challenges from Initial Researches up to the Commercialisation Stage", *Journal of Energy Chemistry*, Vol. 39, pp. 109-143, 2019.

125. Antal, M.J., et al., "Attainment of the Theoretical Yield of Carbon from Biomass", *Industrial & Engineering Chemistry Research*, Vol. 39, No. 11, pp. 4024-4031, 2000.

126. Antal, M.J. and M. Grønli, "The Art, Science, and Technology of Charcoal Production", *Industrial & Engineering Chemistry Research*, Vol. 42, No. 8, pp. 1619-1640, 2003.

127. Fassinou, W.F., et al., "Pyrolysis of Pinus Pinaster in a Two-Stage Gasifier: Influence of Processing Parameters and Thermal Cracking of Tar", *Fuel Processing Technology*, Vol. 90, No. 1, pp. 75-90, 2009.

128. Park, H.J., Y.-K. Park, and J.S. Kim, "Influence of Reaction Conditions and the Char Separation System on the Production of Bio-Oil from Radiata Pine Sawdust by Fast Pyrolysis", *Fuel Processing Technology*, Vol. 89, No. 8, pp. 797-802, 2008.

129. Pattiya, A., *Direct Thermochemical Liquefaction for Energy Applications*, Woodhead Publishing, Duxford, 2018.

130. Kabakcı, S.B. and Ş. Hacıbektaşoğlu, *Pyrolysis*, Intech, 2017.

131. Scheirs, J., *Feedstock Recycling and Pyrolysis of Waste Plastics*, Wiley, West Sussex, 2006.

132. Martínez, J.D., et al., "Co-Pyrolysis of Biomass with Waste Tyres: Upgrading of Liquid Bio-Fuel", *Fuel Processing Technology*, Vol. 119, pp. 263-271, 2014.

133. Abnisa, F. and W.M.A. Wan Daud, "A Review on Co-Pyrolysis of Biomass: An Optional Technique to Obtain a High-Grade Pyrolysis Oil", *Energy Conversion and Management*, Vol. 87, pp. 71-85, 2014.

134. Günay, M.E., L. Türker, and N.A. Tapan, "Significant Parameters and Technological Advancements in Biodiesel Production Systems", *Fuel*, Vol. 250, pp. 27-41, 2019.

135. Sitepu, I.R., et al., "Oleaginous Yeasts for Biodiesel: Current and Future Trends in Biology and Production", *Biotechnology Advances*, Vol. 32, No. 7, pp. 1336-1360, 2014.

136. Ageitos, J.M., et al., "Oily Yeasts as Oleaginous Cell Factories", *Applied Microbiology and Biotechnology*, Vol. 90, No. 4, pp. 1219-1227, 2011.

137. Khot, M., et al., "Lipid Recovery from Oleaginous Yeasts: Perspectives and Challenges for Industrial Applications", *Fuel*, Vol. 259, 2020.

138. Magdouli, S., S.K. Brar, and J.F. Blais, "Morphology and Rheological Behaviour of Yarrowia Lipolytica: Impact of Dissolved Oxygen Level on Cell Growth and Lipid Composition", *Process Biochemistry*, Vol. 65, pp. 1-10, 2018.

139. Darvishi, F., et al., "Yarrowia Lipolytica as a Workhorse for Biofuel Production", *Biochemical Engineering Journal*, Vol. 127, pp. 87-96, 2017.

140. Li, H. and H.S. Alper, "Enabling Xylose Utilization in Yarrowia Lipolytica for Lipid Production", *Biotechnology Journal*, Vol. 11, No. 9, pp. 1230-40, 2016.

141. Yook, S.D., et al., "High-Yield Lipid Production from Lignocellulosic Biomass Using Engineered Xylose-Utilizing Yarrowia Lipolytica", *GCB Bioenergy*, Vol. 12, No. 9, pp. 670-679, 2020.

142. Sarris, D., et al., "Citric Acid, Biomass and Cellular Lipid Production by Yarrowia Lipolytica Strains Cultivated on Olive Mill Wastewater-Based Media", *Journal of Chemical Technology & Biotechnology*, Vol. 86, No. 11, pp. 1439-1448, 2011.

143. Unrean, P., S. Khajeeram, and V. Champreda, "Combining Metabolic Evolution and Systematic Fed-Batch Optimization for Efficient Single-Cell Oil Production from Sugarcane Bagasse", *Renewable Energy*, Vol. 111, pp. 295-306, 2017.

144. Dobrowolski, A., et al., "Efficient Conversion of Crude Glycerol from Various Industrial Wastes into Single Cell Oil by Yeast Yarrowia Lipolytica", *Bioresource Technology*, Vol. 207, pp. 237-243 ,2016.

145. Dragosits, M. and D. Mattanovich, "Adaptive Laboratory Evolution - Principles and Applications for Biotechnology", *Microbial Cell Factories*, Vol. 12, pp. 64, 2013.

146. Daskalaki, A., et al., "Laboratory Evolution Strategies for Improving Lipid Accumulation in Yarrowia Lipolytica", *Applied Microbiology and Biotechnology*, Vol. 103, No. 20, pp. 8585-8596, 2019.

147. Canonico, L., et al., "Conversion of Raw Glycerol to Microbial Lipids by New Metschnikowia and Yarrowia Lipolytica Strains", *Annals of Microbiology*, Vol. 66, No. 4, pp. 1409-1418, 2016.

148. Lopes, M., et al., "Microbial Lipids and Added Value Metabolites Production by Yarrowia Lipolytica from Pork Lard", *Journal of Biotechnology*, Vol. 265, pp. 76-85, 2018.

149. Saran, S., et al., "Process Optimization for Cultivation and Oil Accumulation in an Oleaginous Yeast Rhodosporidium Toruloides A29", *Fuel*, Vol. 188, pp. 324-331, 2017.

150. Papanikolaou, S., et al., "Influence of Glucose and Saturated Free-Fatty Acid Mixtures on Citric Acid and Lipid Production by Yarrowia Lipolytica", *Current Microbiology*, Vol. 52, No. 2, pp. 134-42, 2006.

151. Günay, M.E. and R. Yıldırım, "Recent Advances in Knowledge Discovery for Heterogeneous Catalysis Using Machine Learning", *Catalysis Reviews*, Vol. 63, No. 1, pp. 120-164, 2020.

152. Lamers, D., et al., "Selection of Oleaginous Yeasts for Fatty Acid Production", *BMC Biotechnology*, Vol. 16, No. 1, pp. 45, 2016.

153. Bao, W., et al., "Approaches to Improve the Lipid Synthesis of Oleaginous Yeast Yarrowia Lipolytica: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 149, pp. 111386, 2021.

154. Coşgun, A., M.E. Günay, and R. Yıldırım, "Analysis of Lipid Production from Yarrowia Lipolytica for Renewable Fuel Production by Machine Learning", *Fuel*, Vol. 315, pp. 122817, 2022.

155. Han, D., et al., "Effect of Oxidants on Syngas Synthesis from Biogas over 3 Wt % Ni-Ce-MgO-ZrO2/Al2O3 Catalyst", *Energies*, Vol. 13, No. 2, pp. 297, 2020.

156. Chen, L., P. Gangadharan, and H.H. Lou, "Sustainability Assessment of Combined Steam and Dry Reforming Versus Tri-Reforming of Methane for Syngas

Production", *Asia-Pacific Journal of Chemical Engineering*, Vol. 13, No. 2, pp. 2168, 2018.

157. Yin, W., N. Guilhaume, and Y. Schuurman, "Model Biogas Reforming over Ni-Rh/MgAl2O4 Catalyst: Effect of Gas Impurities", *Chemical Engineering Journal*, Vol. 398, pp. 125534, 2020.

158. Osat, M. and F. Shojaati, "Development of a Trireforming Fluidized-Bed Reactor with a Hydrogen Permselective Membrane", *Chemical Engineering & Technology*, Vol. 45, No. 5, pp. 824-833, 2022.

159. Alli, R.D., et al., "Tri-Reforming of Methane for Syngas Production Using Ni Catalysts: Current Status and Future Outlook", *Catalysis Today*, Vol. 407, pp. 107-124, 2023.

160. Díez-Ramírez, J., et al., "Kinetic, Energetic and Exergetic Approach to the Methane Tri-Reforming Process", International Journal of Hydrogen Energy, Vol. 41, No. 42, pp. 19339-19348, 2016.

161. Song, C. and W. Pan, "Tri-Reforming of Methane: A Novel Concept for Catalytic Production of Industrially Useful Synthesis Gas with Desired H2/CO Ratios", *Catalysis Today*, Vol. 98, No. 4, pp. 463-484, 2004.

162. Lo Faro, M., et al., "Performance Evaluation of a Solid Oxide Fuel Cell Coupled to an External Biogas Tri-Reforming Process", *Fuel Processing Technology*, Vol. 115, pp. 238-245, 2013.

163. Chein, R.-Y. and W.-H. Hsu, "Analysis of Syngas Production from Biogas via the Tri-reforming Process", *Energies*, Vol. 11, No. 5, pp. 1075, 2018.

164. Rahimpour, M.R., Z. Arab Aboosadi, and A.H. Jahanmiri, "Synthesis Gas Production in a Novel Hydrogen and Oxygen Perm-Selective Membranes Tri-

reformer for Methanol Production", *Journal of Natural Gas Science and Engineering*, Vol. 9, pp. 149-159, 2012.

165. Rahnama, H., et al., "Modeling of Synthesis Gas and Hydrogen Production in a Thermally Coupling of Steam and Tri-reforming of Methane with Membranes", *Journal of Industrial and Engineering Chemistry*, Vol. 20, No. 4, pp. 1779-1792, 2014.

166. Manenti, F., et al., "Biogas-fed Solid Oxide Fuel Cell (SOFC) Coupled to Tri-Reforming Process: Modelling and Simulation", *International Journal of Hydrogen Energy*, Vol. 40, No. 42, pp. 14640-14650, 2015.

167. Dwivedi, A., R. Gudi, and P. Biswas, "An Improved Tri-reforming Based Methanol Production Process for Enhanced CO2 Valorization", *International Journal of Hydrogen Energy*, Vol. 42, No. 36, pp. 23227-23241, 2017.

168. Rezaei, E. and L.J.J. Catalan, "Evaluation of CO2 Utilization for Methanol Production via Tri-reforming of Methane", *Journal of CO2 Utilization*, Vol. 42, pp. 101272, 2020.

169. Graciano, J.E.A., et al., "Production of Fuels from CO2-rich Natural Gas using Fischer-Tropsch Synthesis Coupled to Trireforming Process", *Computer Aided Chemical Engineering*, pp. 2659-2664, 2017.

170. Wu, K.T., C.T. Yu, and R.Y. Chein, "Numerical Modeling on Catalytic Tri-reforming Reaction of Methane for Syngas Production", *Energy Procedia*, Vol. 105, pp. 4198-4203, 2017.

171. Zhang, Y., et al., "Thermodynamic Analyses of Tri-reforming Reactions to Produce Syngas", *Energy & Fuels*, Vol. 28, No. 4, pp. 2717-2726, 2014.

172. Chein, R.-Y., C.-Y. Wang, and C.-T. Yu, "Parametric Study on Catalytic Tri-reforming of Methane for Syngas Production", *Energy*, Vol. 118, pp. 1-17, 2017.

173. Alsaffar, M.A., B.V. Ayodele, and S.I. Mustapa, "Scavenging Carbon Deposition on Alumina Supported Cobalt Catalyst During Renewable Hydrogen-rich Syngas Production by Methane Dry Reforming Using Artificial Intelligence Modeling Technique", *Journal of Cleaner Production*, Vol. 247, pp. 119168, 2020.

174. Ayodele, B.V., et al., "Carbon Dioxide Reforming of Methane over Ni-based Catalysts: Modeling the Effect of Process Parameters on Greenhouse Gasses Conversion Using Supervised Machine Learning Algorithms", *Chemical Engineering and Processing - Process Intensification*, Vol. 166, pp. 108484, 2021.

175. Vellayappan, K., et al., "Impacts of Catalyst and Process Parameters on Ni-catalyzed Methane Dry Reforming via Interpretable Machine Learning", *Applied Catalysis B: Environmental*, Vol. 330, pp. 122593, 2023.

176. Şener, A.N., et al., "Statistical Review of Dry Reforming of Methane Literature Using Decision Tree and Artificial Neural Network Analysis", *Catalysis Today*, Vol. 299, pp. 289-302, 2018.

177. Adeniyi, A.G., J.O. Ighalo, and G. Marques, "Utilisation of Machine Learning Algorithms for the Prediction of Syngas Composition from Biomass Bio-oil Steam Reforming", *International Journal of Sustainable Energy*, Vol. 40, No. 4, pp. 310-325, 2021.

178. Lin, S., et al., "Dry Reforming of Methane on Doped Ni Nanoparticles: Feature-assisted Optimizations and Ranking of Doping Metals for Direct Activations of CH4 and CO2", *Nano Research*, Vol. 15, No. 10, pp. 9670-9682, 2022.

179. Baysal, M., M.E. Günay, and R. Yıldırım, "Decision Tree Analysis of Past Publications on Catalytic Steam Reforming to Develop Heuristics for High

Performance: A Statistical Review", *International Journal of Hydrogen Energy*, Vol. 42, No. 1, pp. 243-254, 2017.

180. Lee, J., et al., "Machine Learning-based Energy Optimization for On-site SMR Hydrogen Production", *Energy Conversion and Management*, Vol. 244, pp. 114438, 2021.

181. Hong, S., et al., "Multi-objective Optimization of CO2 Emission and Thermal Efficiency for On-site Steam Methane Reforming Hydrogen Production Process Using Machine Learning", *Journal of Cleaner Production*, Vol. 359, pp. 132133, 2022.

182. Odabaşı, Ç., M.E. Günay, and R. Yıldırım, "Knowledge Extraction for Water Gas Shift Reaction Over Noble Metal Catalysts from Publications in the Literature Between 2002 and 2012", *International Journal of Hydrogen Energy*, Vol. 39, No. 11, pp. 5733-5746, 2014.

183. Gul, H., M.Y. Arshad, and M.W. Tahir, "Production of H2 via Sorption Enhanced Auto-thermal Reforming for Small Scale Applications - A Process Modeling and Machine Learning Study", *International Journal of Hydrogen Energy*, Vol. 48, No. 34, pp. 12622-12635, 2023.

184. Ugwu, L.I., Y. Morgan, and H. Ibrahim, "Application of Density Functional Theory and Machine Learning in Heterogenous-Based Catalytic Reactions for Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 47, No. 4, pp. 2245-2267, 2022.

185. Deng, H. and Y. Guo, "Artificial Neural Network Model for the Prediction of Methane Bi-reforming Products Using CO2 and Steam", *Processes*, Vol. 10, No. 6, pp. 1052, 2022.

186. Larose, D.T. and C.D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edition, Wiley, New Jersey, 2014.

187. Erdem Günay, M. and R. Yıldırım, "Recent Advances in Knowledge Discovery for Heterogeneous Catalysis Using Machine Learning", *Catalysis Reviews*, Vol. 63, No. 1, pp. 120-164, 2020.

188. Jain, A., et al., "Commentary: The Materials Project: A materials Genome Approach to Accelerating Materials Innovation", *APL Materials*, Vol. 1, No. 1, 2013.

189. Alpaydin, E., *Introduction to Machine Learning*, Fourth Edition, The MIT Press, Cambridge, 2020.

190. Awad, M. and R. Khanna, *Theories, Concepts, and Applications for Engineers and System Designers*, Apress, Berkeley, 2015.

191. Vapnik, V.N., *The Nature of Statistical Learning Theory*, Second Edition, Springer, New York, 2000.

192. Jain, A., et al., "New Opportunities for Materials Informatics: Resources and Data Mining Techniques for Uncovering Hidden Relationships", *Journal of Materials Research*, Vol. 31, No. 8, pp. 977-994, 2016.

193. Kirklin, S., et al., "The Open Quantum Materials Database (OQMD): Assessing the Accuracy ff DFT Formation Energies", *npj Computational Materials*, Vol. 1, No. 1, pp. 1-15, 2015.

194. Curtarolo, S., et al., "Aflowlib.org: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations", *Computational Materials Science*, Vol. 58, pp. 227-235, 2012.

195. Landis, D.D., et al., "The Computational Materials Repository", *Computing in Science & Engineering*, Vol. 14, No. 6, pp. 51-57, 2012.

196. Huber, S.P., et al., "AiiDA 1.0, A Scalable Computational Infrastructure for Automated Reproducible Workflows and Data Provenance", *Scientific data*, Vol. 7, No. 1, pp. 1-18, 2020.

197. Zhang, Q., et al., "Firework: Data Processing and Sharing for Hybrid Cloud-Edge Analytics", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 29, No. 9, pp. 2004-2017, 2018.

198. Mathew, K., et al., "Atomate: A High-Level Interface to Generate, Execute, and Analyze Computational Materials Science Workflows", *Computational Materials Science*, Vol. 139, p. 140-152, 2017.

199. Ong, S.P., et al., "Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis", *Computational Materials Science*, Vol. 68, pp. 314-319, 2013.

200. Zhang, Y. and C. Ling, "A Strategy to Apply Machine Learning to Small Datasets in Materials Science", *npj Computational Materials*, Vol. 4, No. 1, 2018.

201. Altikat, A. and M.H. Alma, "Prediction Carbonization Yields and the Sensitivity Analyses Using Deep Learning Neural Networks and Support Vector Machines", *International Journal of Environmental Science and Technology*, Vol. 20, No. 5, pp. 5071-5080, 2022.

202. Kaya, M. and S. Hajimirza, "Using a Novel Transfer Learning Method for Designing Thin Film Solar Cells with Enhanced Quantum Efficiencies", *Scientific Reports*, Vol. 9, No. 1, pp. 5034, 2019.

203. Liu, R., et al., "Enabling Data-Limited Chemical Bioactivity Predictions Through Deep Neural Network Transfer Learning", *Journal of Computer-Aided Molecular Design*, Vol. 36, No. 12, pp. 867-878, 2022.

204. Cohen, S., E. Ruppin, and G. Dror, "Feature Selection Based on the Shapley Value", *Nineteenth International Joint Conference on Artificial Intelligence,* Edinburgh, Scotland, UK, 2005.

205. Molnar, C., *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*, Second Edition, Independently published, 2022.

206. Feng, S., H. Zhou, and H. Dong, "Using Deep Neural Network with Small Dataset to Predict Material Defects"*, Materials & Design*, Vol. 162, pp. 300-310, 2019.

207. Ma, R., Y.J. Colon, and T. Luo, "Transfer Learning Study of Gas Adsorption in Metal-Organic Frameworks", *ACS Applied Materials & Interfaces Journal*, Vol. 12, No. 30, pp. 34041-34048, 2020.

208. Yaoyang, X. and W.J. Boeing, "Mapping Biofuel Field: A Bibliometric Evaluation of Research Output", *Renewable and Sustainable Energy Reviews*, Vol. 28, pp. 82-91, 2013.

209. Kuo, E.Y., et al., "Multiomics Approaches and Genetic Engineering of Metabolism for Improved Biorefinery and Wastewater Treatment in Microalgae", *Biotechnology Journal*, Vol. 17, No. 8, pp. e2100603, 2022.

210. Azari, A., et al., "Predictive Model of Algal Biofuel Production Based on Experimental Data", *Algal Research*, Vol. 47, 2020.

211. Vatcheva, I., et al., "Experiment Selection for the Discrimination of Semi-Quantitative Models of Dynamical Systems", *Artificial Intelligence*, Vol. 170, No. 4-5, pp. 472-506, 2006.

212. Del Rio-Chanona, E.A., et al., "Comparison of Physics-Based and Data-Driven Modelling Techniques for Dynamic Optimisation of Fed-Batch Bioprocesses", *Biotechnology & Bioengineering*, Vol. 116, No. 11, pp. 2971-2982, 2019.

213. Aghbashlo, M., et al., "Machine Learning Technology in Biodiesel Research: A Review", *Progress in Energy and Combustion Science*, Vol. 85, 2021.

214. Ascher, S., I. Watson, and S. You, "Machine Learning Methods for Modelling the Gasification and Pyrolysis of Biomass and Waste", *Renewable and Sustainable Energy Reviews*, Vol. 155, 2022.

215. Ning, H., R. Li, and T. Zhou, "Machine Learning for Microalgae Detection and Utilization", *Frontiers in Marine Science*, Vol. 9, 2022.

216. Sundui, B., et al., "Applications of Machine Learning Algorithms for Biological Wastewater Treatment: Updates and Perspectives", *Clean Technologies and Environmental Policy*, Vol. 23, No. 1, pp. 127-143, 2021.

217. Condor, B.E., et al., "Optimization and Modeling of Carbohydrate Production in Microalgae for Use as Feedstock in Bioethanol Fermentation", *International Journal of Energy Research*, Vol. 46, No. 13, pp. 19300-19312, 2022.

218. Hossain, S.M.Z., et al., "Soft-Computing Modeling and Multiresponse Optimization for Nutrient Removal Process from Municipal Wastewater Using Microalgae", *Journal of Water Process Engineering*, Vol. 45, 2022.

219. Srivastava, G., A.K. Paul, and V.V. Goud, "Optimization of Non-Catalytic Transesterification of Microalgae Oil to Biodiesel Under Supercritical Methanol Condition", *Energy Conversion and Management*, Vol. 156, pp. 269-278, 2018.

220. Desai, K.M., et al., "Comparison of Artificial Neural Network (ANN) and Response Surface Methodology (RSM) in Fermentation Media Optimization: Case Study of

Fermentative Production of Scleroglucan", *Biochemical Engineering Journal*, Vol. 41, No. 3, pp. 266-273, 2008.

221. Zhang, W., et al., "Machine Learning Prediction and Optimization of Bio-Oil Production from Hydrothermal Liquefaction of Algae", *Bioresource Technology*, Vol. 342, pp. 126011, 2021.

222. Panahi, B., et al., "Integration of Cross Species RNA-Seq Meta-Analysis and Machine-Learning Models Identifies rhe Most Important Salt Stress-Responsive Pathways in Microalga Dunaliella", *Frontiers in Genetics*, Vol. 10, pp. 752, 2019.

223. Pozzobon, V., et al., "Machine Learning Processing of Microalgae Flow Cytometry Readings: Illustrated with Chlorella Vulgaris Viability Assays", *Journal of Applied Phycology*, Vol. 32, No. 5, pp. 2967-2976, 2020.

224. Reimann, R., et al., "Classification of Dead and Living Microalgae Chlorella Vulgaris by Bioimage Informatics and Machine Learning", *Algal Research*, Vol. 48, 2020.

225. Wang, Y., et al., "Identification of Living and Dead Microalgae Cells with Digital Holography and Verified in The East China Sea", *Marine Pollution Bulletin*, Vol. 163, pp. 111927, 2021.

226. Dhindsa, A., et al., "An Improvised Machine Learning Model Based on Mutual Information Feature Selection Approach for Microbes Classification", *Entropy*, Vol. 23, No. 2, 2021.

227. Sonmez, M.E., et al., "Convolutional Neural Network - Support Vector Machine Based Approach for Classification of Cyanobacteria and Chlorophyta Microalgae Groups", *Algal Research*, Vol. 61, 2022.

228. Salmi, P., et al., "Assessment of Microalgae Species, Biomass, and Distribution from Spectral Images Using a Convolution Neural Network", *Journal of Applied Phycology*, Vol. 34, No. 3, pp. 1565-1575, 2022.

229. Xu, Z., et al., "Classification, Identification, and Growth Stage Estimation of Microalgae Based on Transmission Hyperspectral Microscopic Imaging and Machine Learning", *Optics Express*, Vol. 28, No. 21, pp. 30686-30700, 2020.

230. Supriyanto, et al., "Artificial Neural Networks Model for Estimating Growth of Polyculture Microalgae in an Open Raceway Pond", *Biosystems Engineering*, Vol. 177, pp. 122-129, 2019.

231. Ching, P.M.L., et al., "Early Prediction of Spirulina Platensis Biomass Yield for Biofuel Production Using Machine Learning", *Clean Technologies and Environmental Policy*, Vol. 24, No. 7, pp. 2283-2293, 2022.

232. Doan, Y.T.-T., et al., "Optimization of Spirulina Sp. Cultivation Using Reinforcement Learning with State Prediction Based on LSTM Neural Network", *Journal of Applied Phycology*, Vol. 33, No. 5, pp. 2733-2744, 2021.

233. Singh, V. and V. Mishra, "Exploring the Effects of Different Combinations of Predictor Variables for the Treatment of Wastewater by Microalgae and Biomass Production", *Biochemical Engineering Journal*, Vol. 174, 2021.

234. Singh, V. and V. Mishra, "Evaluation of the Effects of Input Variables on the Growth of Two Microalgae Classes During Wastewater Treatment", *Water Research*, Vol. 213, pp. 118165, 2022.

235. García Nieto, P.J., et al., "A Hybrid PSO Optimized SVM-Based Model for Predicting a Successful Growth Cycle of the Spirulina Platensis from Raceway Experiments Data", *Journal of Computational and Applied Mathematics*, Vol. 291, pp. 293-303, 2016.

236. Guo, Y., et al., "Effects of Microplastics on Growth, Phenanthrene Stress, and Lipid Accumulation in a Diatom, Phaeodactylum Tricornutum", *Environmental Pollution*, Vol. 257, pp. 113628, 2020.

237. Long, B., et al., "Machine Learning-Informed and Synthetic Biology-Enabled Semi-Continuous Algal Cultivation to Unleash Renewable Fuel Productivity", *Nature Communications*, Vol. 13, No. 1, pp. 541, 2022.

238. Chen, C.-L., J.-S. Chang, and D.-J. Lee, "Dewatering and Drying Methods for Microalgae", *Drying Technology*, Vol. 33, No. 4, pp. 443-454, 2015.

239. Suparmaniam, U., et al., "Valorization of Fish Bone Waste as Novel Bioflocculant for Rapid Microalgae Harvesting: Experimental Evaluation and Modelling Using Back Propagation Artificial Neural Network", *Journal of Water Process Engineering*, Vol. 47, 2022.

240. Ching, P.M.L., et al., "AI Methods for Modeling the Vacuum Drying Characteristics of Chlorococcum Infusionum for Algal Biofuel Production", *Process Integration and Optimization for Sustainability*, Vol. 5, No. 2, pp. 247-256, 2021.

241. Sultana, N., et al., "Prediction of Biodiesel Production from Microalgal Oil Using Bayesian Optimization Algorithm-Based Machine Learning Approaches", *Fuel*, Vol. 309, 2022.

242. Salam, S. and T.N. Verma, "Appending Empirical Modelling to Numerical Solution for Behaviour Characterisation of Microalgae Biodiesel", *Energy Conversion and Management*, Vol. 180, pp. 496-510, 2019.

243. Suvarna, M., et al., "Predicting Biodiesel Properties and its Optimal Fatty Acid Profile via Explainable Machine Learning", *Renewable Energy*, Vol. 189, pp. 245-258, 2022.

244. Correa, I., et al., "Deep Learning for Microalgae Classification", *16th IEEE International Conference on Machine Learning and Applications*, pp. 20-25, 2017.

245. Rodríguez-Rángel, H., et al., "Machine Learning Methods Modeling Carbohydrate-Enriched Cyanobacteria Biomass Production in Wastewater Treatment Systems", *Energies*, Vol. 15, No. 7, 2022.

246. Bong, J.T., et al., "Artificial Neural Network Approach for Co-Pyrolysis of Chlorella Vulgaris and Peanut Shell Binary Mixtures Using Microalgae Ash Catalyst", *Energy*, Vol. 207, 2020.

247. Pathy, A., S. Meher, and B. P, "Predicting Algal Biochar Yield Using Extreme Gradient Boosting (XGB) Algorithm of Machine Learning Methods", *Algal Research*, Vol. 50, 2020.

248. Wang, S., et al., "A Machine Learning Model to Predict the Pyrolytic Kinetics of Different Types of Feedstocks", *Energy Conversion and Management*, Vol. 260, 2022.

249. Li, J., et al., "Machine Learning Aided Bio-Oil Production with High Energy Recovery and Low Nitrogen Content from Hydrothermal Liquefaction of Biomass with Experiment Verification", *Chemical Engineering Journal*, Vol. 425, 2021.

250. Cheng, F. and H. Luo, "Evaluating the Minimum Fuel Selling Price of Algae-Derived Biofuel from Hydrothermal Liquefaction", *Bioresource Technology Reports*, Vol. 17, 2022.

251. Katongtung, T., T. Onsree, and N. Tippayawong, "Machine Learning Prediction of Biocrude Yields and Higher Heating Values from Hydrothermal Liquefaction of Wet Biomass and Wastes", *Bioresource Technology*, Vol. 344, pp. 126278, 2022.

252. Cheng, F., et al., "Accuracy of Predictions Made by Machine Learned Models for Biocrude Yields Obtained from Hydrothermal Liquefaction of Organic Wastes", *Chemical Engineering Journal*, Vol. 442, 2022.

253. Bharathiraja, B., et al., "Aquatic Biomass (Algae) as a Future Feed Stock for Bio-Refineries: A Review on Cultivation, Processing and Products", *Renewable and Sustainable Energy Reviews*, Vol. 47, pp. 634-653, 2015.

254. Bryant, H.L., et al., "The Value Oo Post-Extracted Algae Residue", *Algal Research*, Vol. 1, No. 2, pp. 185-193, 2012.

255. Huang, H.-j. and X.-z. Yuan, "Recent Progress in the Direct Liquefaction of Typical Biomass", *Progress in Energy and Combustion Science*, Vol. 49, pp. 59-80, 2015.

256. Rastogi, R.P., et al., "Algal Green Energy - R&D and Technological Perspectives for Biodiesel Production", *Renewable and Sustainable Energy Reviews*, Vol. 82, pp. 2946-2969, 2018.

257. Khenkin, M.V., et al., "Consensus Statement for Stability Assessment and Reporting for Perovskite Photovoltaics Based on ISOS Procedures", *Nature Energy*, Vol. 5, No. 1, pp. 35-49, 2020.

258. Pradhan, D., S. Jaiswal, and A.K. Jaiswal, "Artificial Neural Networks in Valorization Process Modeling of Lignocellulosic Biomass", *Biofuels, Bioproducts and Biorefining*, Vol. 16, No. 6, pp. 1849-1868, 2022.

259. Sewsynker-Sukai, Y., F. Faloye, and E.B.G. Kana, "Artificial Neural Networks: An Efficient Tool for Modelling and Optimization of Biofuel Production (A Mini Review)", *Biotechnology & Biotechnological Equipment*, Vol. 31, No. 2, pp. 221-235, 2016.

260. Li, H., et al., "Machine-Learning-Aided Thermochemical Treatment of Biomass: A Review", *Biofuel Research Journal*, Vol. 10, No. 1, pp. 1786-1809, 2023.

261. Li, J., et al., "A Review of Computational Modeling Techniques for Wet Waste Valorization: Research Trends and Future Perspectives", *Journal of Cleaner Production*, Vol. 367, 2022.

262. Guo, H.N., et al., "Application of Machine Learning Methods for the Prediction of Organic Solid Waste Treatment and Recycling Processes: A Review", *Bioresource Technology*, Vol. 319, pp. 124114, 2021.

263. Ge, S., et al., "Progress in Pyrolysis Conversion of Waste into Value-Added Liquid Pyro-Oil, with Focus on Heating Source and Machine Learning Analysis", *Energy Conversion and Management*, Vol. 245, 2021.

264. Phromphithak, S., T. Onsree, and N. Tippayawong, "Machine Learning Prediction of Cellulose-Rich Materials from Biomass Pretreatment with Ionic Liquid Solvents", *Bioresource Technology*, Vol. 323, pp. 124642, 2021.

265. Parkhey, P., et al., "Artificial Neural Network And Response Surface Methodology: A Comparative Analysis for Optimizing Rice Straw Pretreatment and Saccharification", *Preparative Biochemistry & Biotechnology*, Vol. 50, No. 8, pp. 768-780, 2020.

266. Aruwajoye, G.S., et al., "Intelligent Modelling of Fermentable Sugar Concentration and Combined Severity Factor (CSF) Index from Pretreated Starch-Based Lignocellulosic Biomass", *Biomass Conversion and Biorefinery*, 2022.

267. Lee, K.M., et al., "Synergistic Ultrasound-Assisted Organosolv Pretreatment of Oil Palm Empty Fruit Bunches for Enhanced Enzymatic Saccharification: An Optimization Study Using Artificial Neural Networks", *Biomass and Bioenergy*, Vol. 139, 2020.

268. Moodley, P., D.C.S. Rorke, and E.B. Gueguim Kana, "Development of Artificial Neural Network Tools for Predicting Sugar Yields from Inorganic Salt-Based Pretreatment of Lignocellulosic Biomass", *Bioresource Technology*, Vol. 273, pp. 682-686, 2019.

269. Onay, M., "Sequential Modelling for Carbohydrate and Bioethanol Production from Chlorella Saccharophila CCALA 258: A Complementary Experimental and Theoretical Approach for Microalgal Bioethanol Production", *Environmental Science and Pollution Research*, Vol. 29, No. 10, pp. 14316-14332, 2022.

270. Chen, W.-H., et al., "Forecast of Glucose Production from Biomass Wet Torrefaction Using Statistical Approach Along with Multivariate Adaptive Regression Splines, Neural Network and Decision Tree", *Applied Energy*, Vol. 324, 2022.

271. Gitifar, V., R. Eslamloueyan, and M. Sarshar, "Experimental Study and Neural Network Modeling of Sugarcane Bagasse Pretreatment with H2SO4 and O3 for Cellulosic Material Conversion to Sugar", *Bioresource Technology*, Vol. 148, pp. 47-52, 2013.

272. Plazas Tovar, L., et al., "Prediction of Overall Glucose Yield in Hydrolysis of Pretreated Sugarcane Bagasse Using a Single Artificial Neural Network: Good Insight for Process Development", *Journal of Chemical Technology & Biotechnology*, Vol. 93, No. 4, pp. 1031-1043, 2018.

273. Ethaib, S., et al., "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 Computing and Applications*, Vol. 30, No. 4, pp. 1111-1121, 2016.

274. Chang, C.-W., et al., "A Study on the Enzymatic Hydrolysis of Steam Exploded Napiergrass with Alkaline Treatment Using Artificial Neural Networks and

Regression Analysis", *Journal of the Taiwan Institute of Chemical Engineers*, Vol. 42, No. 6, pp. 889-894, 2011.

275. Rivera, E.C., et al., "Enzymatic Hydrolysis of Sugarcane Bagasse for Bioethanol Production: Determining Optimal Enzyme Loading Using Neural Networks", *Journal of Chemical Technology & Biotechnology*, Vol. 85, No. 7, pp. 983-992, 2010.

276. Gama, R., et al., "Using an Artificial Neural Network to Predict the Optimal Conditions for Enzymatic Hydrolysis of Apple Pomace", *3 Biotech*, Vol. 7, No. 2, pp. 138, 2017.

277. Giordano, P.C., et al., "Optimization of the Hydrolysis of Lignocellulosic Residues by Using Radial Basis Functions Modeling and Particle Swarm Optimization", *Biochemical Engineering Journal*, Vol. 80, pp. 1-9, 2013.

278. Ganguly, P. and P. Das, "Integral Approach for Second-Generation Bio-Ethanol Production and Wastewater Treatment Using Peanut Shell Waste: Yield, Removal, and ANN Studies", *Biomass Conversion and Biorefinery*, Vol. 14, pp. 689-699, 2022.

279. Chouaibi, M., et al., "Production of Bioethanol from Pumpkin Peel Wastes: Comparison Between Response Surface Methodology (RSM) and Artificial Neural Networks (ANN)", *Industrial Crops and Products*, Vol. 155, 2020.

280. Shet, V.B., et al., "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*, Vol. 8, No. 2, pp. 127, 2018.

281. B. Shet, V., et al., "Optimization of Microwave Assisted H2SO4 Hydrolysis of Cocoa Pod Shells: Comparison Between Response Surface Methodology and

Artificial Neural Network and Production of Bioethanol Thereof", *Journal of Microbiology, Biotechnology and Food Sciences*, Vol. 7, No. 5, pp. 473-477, 2018.

282. Mondal, P., et al., "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*, Vol. 11, No. 1, pp. 28, 2021.

283. Sebayang, A.H., et al., "Optimization of Bioethanol Production from Sorghum Grains Using Artificial Neural Networks Integrated with Ant Colony", *Industrial Crops and Products*, Vol. 97, pp. 146-155, 2017.

284. Suresh, T., et al., "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 and Bioenergy*, Vol. 142, 2020.

285. Vinitha, N., J. Vasudevan, and K.P. Gopinath, "Bioethanol Production Optimization Through Machine Learning Algorithm Approach: Biomass Characteristics, Saccharification, and Fermentation Conditions for Enzymatic Hydrolysis", *Biomass Conversion and Biorefinery*, Vol. 13, pp. 7287- 7299, 2022.

286. Fischer, J., et al., "Machine Learning Techniques Applied to Lignocellulosic Ethanol in Simultaneous Hydrolysis and Fermentation", *Brazilian Journal of Chemical Engineering*, Vol. 34, No. 1, pp. 53-63, 2017.

287. Das, S., et al., "Artificial Neural Network Modelling of Xylose Yield from Water Hyacinth by Dilute Sulphuric Acid Hydrolysis for Ethanol Production", *International Journal of Environmental Technology and Management*, Vol. 19, No. 2, 2016.

288. Dharmalingam, B., et al., "Organic Acid Pretreatments and Optimization Techniques for Mixed Vegetable Waste Biomass Conversion into Biofuel Production", *BioEnergy Research*, Vol.16, pp. 1667-1682, 2022.

289. Mohaptra, S., et al., "Optimization of Delignification of Two Pennisetum Grass Species by NaOH Pretreatment Using Taguchi and ANN Statistical Approach", *Environmental Technology*, Vol. 37, No. 8, pp. 940-51, 2016.

290. Rashid, T., et al., "Enhanced Lignin Extraction and Optimisation from Oil Palm Biomass Using Neural Network Modelling", *Fuel*, Vol. 293, 2021.

291. Singhal, A., et al., "Pretreatment of Leucaena Leucocephala Wood by Acidified Glycerol: Optimization, Severity Index and Correlation Analysis", *Bioresource Technology*, Vol. 265, pp. 214-223, 2018.

292. Haldar, D., A.M. Shabbirahmed, and B. Mahanty, "Multivariate Regression and Artificial Neural Network Modelling of Sugar Yields from Acid Pretreatment and Enzymatic Hydrolysis of Lignocellulosic Biomass", *Bioresource Technology*, Vol. 370, pp. 128519, 2023.

293. Luo, H., et al., "Prediction of Phenolic Compounds and Glucose Content from Dilute Inorganic Acid Pretreatment of Lignocellulosic Biomass Using Artificial Neural Network Modeling", *Bioresources and Bioprocessing*, Vol. 8, No. 1, 2021.

294. Dave, N., et al., "Modelling of Fermentative Bioethanol Production from Indigenous Ulva Prolifera Biomass by Saccharomyces Cerevisiae NFCCI1248 Using an Integrated ANN-GA Approach", *Science of the Total Environment*, Vol. 791, pp. 148429, 2021.

295. Grahovac, J., et al., "Modelling and Prediction of Bioethanol Production from Intermediates and Byproduct of Sugar Beet Processing Using Neural Networks", *Renewable Energy*, Vol. 85, pp. 953-958, 2016.

296. Konishi, M., "Bioethanol Production Estimated from Volatile Compositions in Hydrolysates of Lignocellulosic Biomass by Deep Learning", *Journal of Bioscience and Bioengineering*, Vol. 129, No. 6, pp. 723-729, 2020.

297. Pereira, R.D., A.C. Badino, and A.J.G. Cruz, "Framework Based on Artificial Intelligence to Increase Industrial Bioethanol Production", *Energy & Fuels*, Vol. 34, No. 4, pp. 4670-4677, 2020.

298. Taiwo, A., T. Madzimbamuto, and T. Ojumu, "Optimization of Corn Steep Liquor Dosage and Other Fermentation Parameters for Ethanol Production by Saccharomyces cerevisiae Type 1 and Anchor Instant Yeast", *Energies*, Vol. 11, No. 7, 2018.

299. Selvakumar, P., S. Kavitha, and P. Sivashanmugam, "Optimization of Process Parameters for Efficient Bioconversion of Thermo-chemo Pretreated Manihot Esculenta Crantz YTP1 Stem to Ethanol", *Waste and Biomass Valorization*, Vol. 10, No. 8, pp. 2177-2191, 2018.

300. Ezzatzadegan, L., et al., "Experimental and Artificial Intelligence Modelling Study of Oil Palm Trunk Sap Fermentation", *Energies*, Vol. 14, No. 8, 2021.

301. Betiku, E. and A.E. Taiwo, "Modeling and Optimization of Bioethanol Production from Breadfruit Starch Hydrolyzate vis-a-vis Response Surface Methodology and Artificial Neural Network", *Renewable Energy*, Vol. 74, pp. 87-94, 2015.

302. Raj, T., et al., "Recent Advances in Commercial Biorefineries for Lignocellulosic Ethanol Production: Current Status, Challenges and Future Perspectives", *Bioresource Technology*, Vol. 344, pp. 126292, 2022.

303. Campion, L., et al., "The Costs and Benefits of Biochar Production and Use: A Systematic Review", *Journal of Cleaner Production*, Vol. 408, 2023.

304. Helleckes, L.M., et al., "Machine Learning in Bioprocess Development: From Promise to Practice", *Trends in Biotechnology*, Vol. 41, No. 6, pp. 817-835, 2023.

305. Hai, A., et al., "Machine Learning Models for the Prediction of Total Yield and Specific Surface Area of Biochar Derived from Agricultural Biomass by Pyrolysis", *Environmental Technology & Innovation*, Vol. 30, 2023.

306. Li, J., et al., "Multi-Task Prediction and Optimization of Hydrochar Properties from High-Moisture Municipal Solid Waste: Application of Machine Learning on Waste-to-Resource", *Journal of Cleaner Production*, Vol. 278, 2021.

307. Mari Selvam, S. and P. Balasubramanian, "Influence of Biomass Composition and Microwave Pyrolysis Conditions on Biochar Yield and its Properties: A Machine Learning Approach", *BioEnergy Research*, Vol. 16, No. 1, pp. 138-150, 2022.

308. Zhu, X., et al., "Machine Learning-Assisted Exploration for Carbon Neutrality Potential of Municipal Sludge Recycling via Hydrothermal Carbonization", *Bioresource Technology*, Vol. 369, pp. 128454, 2023.

309. Merdun, H. and I.V. Sezgin, "Modeling of Pyrolysis Product Yields by Artificial Neural Networks", *International Journal of Renewable Energy Research*, Vol. 8, No. 2, 2018.

310. Tsekos, C., S. Tandurella, and W. de Jong, "Estimation of Lignocellulosic Biomass Pyrolysis Product Yields Using Artificial Neural Networks", *Journal of Analytical and Applied Pyrolysis*, Vol. 157, 2021.

311. Li, J., et al., "Fuel Properties of Hydrochar and Pyrochar: Prediction and Exploration with Machine Learning", *Applied Energy*, Vol. 269, 2020.

312. Ewees, A.A. and M.A. Elaziz, "Improved Adaptive Neuro-Fuzzy Inference System Using Gray Wolf Optimization: A Case Study in Predicting Biochar Yield", *Journal of Intelligent Systems*, Vol. 29, No. 1, pp. 924-940, 2019.

313. Haq, Z.U., et al., "Comparative Study of Machine Learning Methods Integrated with Genetic Algorithm and Particle Swarm Optimization for Bio-Char Yield Prediction", *Bioresource Technology*, Vol. 363, pp. 128008, 2022.

314. Khan, M., et al., "Artificial Neural Networks for the Prediction of Biochar Yield: A Comparative Study of Metaheuristic Algorithms", *Bioresource Technol*ogy, Vol. 355, pp. 127215, 2022.

315. Leng, L., et al., "Machine Learning Predicting and Engineering the Yield, N Content, and Specific Surface Area of Biochar Derived From Pyrolysis Of Biomass", *Biochar*, Vol. 4, No. 1, 2022.

316. Li, H., et al., "Machine Learning Assisted Predicting and Engineering Specific Surface Area and Total Pore Volume of Biochar", *Bioresource Technol*ogy, Vol. 369, pp. 128417, 2023.

317. Leng, E., et al., "Prediction of Three-Phase Product Distribution and Bio-Oil Heating Value of Biomass Fast Pyrolysis Based on Machine Learning", *Energy*, Vo. 236, 2021.

318. Cheng, F., H. Luo, and L.M. Colosi, "Slow Pyrolysis as a Platform for Negative Emissions Technology: An Integration of Machine Learning Models, Life Cycle Assessment, and Economic Analysis", *Energy Conversion and Management*, Vol. 223, 2020.

319. Li, T.Y., et al., "Prediction of Char Production from Slow Pyrolysis of Lignocellulosic Biomass Using Multiple Nonlinear Regression and Artificial Neural Network", *Journal of Analytical and Applied Pyrolysis*, Vol. 159, 2021.

320. Aydinli, B., et al., "The Prediction of Potential Energy and Matter Production from Biomass Pyrolysis with Artificial Neural Network", *Energy Exploration & Exploitation*, Vol. 35, No. 6, pp. 698-712, 2017.

321. Cao, H., Y. Xin, and Q. Yuan, "Prediction of Biochar Yield from Cattle Manure Pyrolysis via Least Squares Support Vector Machine Intelligent Approach", *Bioresource Technology*, Vol. 202, pp. 158-164, 2016.

322. Gu, C., et al., "Prediction of Gas-Liquid-Solid Product Distribution After Solid Waste Pyrolysis Process Based on Artificial Neural Network Model", *International Journal of Energy Research*, Vol. 45, No. 9, pp. 13786-13800, 2021.

323. Tang, J.Y., et al., "Prediction Model for Biochar Energy Potential Based on Biomass Properties and Pyrolysis Conditions Derived from Rough Set Machine Learning", *Environmental Technology*, Vol. 45, No. 15, p. 1-15, 2023.

324. Li, Y., R. Gupta, and S. You, "Machine Learning Assisted Prediction of Biochar Yield and Composition via Pyrolysis Of Biomass", *Bioresource Technology*, Vol. 359, pp. 127511, 2022.

325. Yang, Y., et al., "Biomass Microwave Pyrolysis Characterization by Machine Learning for Sustainable Rural Biorefineries", *Renewable Energy*, Vol. 201, pp. 70-86, 2022.

326. Chen, J., et al., "The Estimation of the Higher Heating Value of Biochar by Data-Driven Modeling", *Journal of Renewable Materials*, Vol. 10, No. 6, pp. 1555-1574, 2022.

327. Dong, Z., et al., "Machine Learning Prediction of Pyrolytic Products of Lignocellulosic Biomass Based on Physicochemical Characteristics and Pyrolysis Conditions", *Bioresource Technology*, Vol. 367, pp. 128182, 2023.

328. Chen, W.-H., et al., "A Comparative Analysis of Biomass Torrefaction Severity Index Prediction from Machine Learning", *Applied Energy*, Vol. 324, 2022.

329. Kardani, N., et al., "Machine Learning Prediction of the Conversion of Lignocellulosic Biomass During Hydrothermal Carbonization", *Biofuels*, Vol. 13, No. 6, pp. 703-715, 2021.

330. Mu, L., et al., "Prediction and Evaluation of Fuel Properties of Hydrochar from Waste Solid Biomass: Machine Learning Algorithm Based on Proposed PSO–NN Model", *Fuel*, Vol. 318, 2022.

331. Cheng, F., M.D. Porter, and L.M. Colosi, "Is Hydrothermal Treatment Coupled with Carbon Capture and Storage an Energy-Producing Negative Emissions Technology?", *Energy Conversion and Management*, Vol. 203, 2020.

332. Djandja, O.S., et al., "A Novel Machine Learning-Based Approach for Prediction of Nitrogen Content in Hydrochar from Hydrothermal Carbonization of Sewage Sludge", *Energy*, Vol. 232, 2021.

333. Djandja, O.S., et al., "Random Forest-Based Modeling for Insights on Phosphorus Content in Hydrochar Produced from Hydrothermal Carbonization of Sewage Sludge", *Energy*, Vol. 245, 2022.

334. Li, L., et al., "Quantifying the Sensitivity of Feedstock Properties and Process Conditions on Hydrochar Yield, Carbon Content, and Energy Content", *Bioresource Technology*, Vol. 262, pp. 284-293, 2018.

335. García Nieto, P.J., et al., "Forecast of the Higher Heating Value in Biomass Torrefaction by Means of Machine Learning Techniques", *Journal of Computational and Applied Mathematics*, Vol. 357, pp. 284-301, 2019.

336. Onsree, T. and N. Tippayawong, "Machine Learning Application to Predict Yields of Solid Products from Biomass Torrefaction", *Renewable Energy*, Vol. 167, pp. 425-432, 2021.

337. Onsree, T., et al., "Interpretable Machine-Learning Model with a Collaborative Game Approach to Predict Yields and Higher Heating Value of Torrefied Biomass", *Energy*, Vol. 249, 2022.

338. Shafizadeh, A., et al., "Machine Learning Predicts and Optimizes Hydrothermal Liquefaction of Biomass", *Chemical Engineering Journal*, Vol. 445, 2022.

339. Hamid, Y., et al., "Functionalized Biochars: Synthesis, Characterization, and Applications for Removing Trace Elements from Water", *Journal of Hazardous Materials*, Vol. 437, pp. 129337, 2022.

340. Andre, A., et al., "Biotechnological Conversions of Bio-Diesel-Derived Crude Glycerol by Yarrowia Lipolytica Strains", *Engineering in Life Sciences*, Vol. 9, No. 6, pp. 468-478, 2009.

341. Chatzifragkou, A., et al., "Biotechnological Conversions of Biodiesel Derived Waste Glycerol by Yeast and Fungal Species", *Energy*, Vol. 36, No. 2, pp. 1097-1108, 2011.

342. Enshaeieh, M., A. Abdoli, and I. Nahvi, "Medium Optimization for Biotechnological Production of Single Cell Oil Using Yarrowia Lipolytica M7 and Candida Sp", *Journal of Cell and Molecular Research*, Vol. 5, No. 1, pp. 17-23, 2013.

343. Papanikolaou, S., et al., "Importance of the Methyl-Citrate Cycle on Glycerol Metabolism in the Yeast Yarrowia Lipolytica", *Journal of Biotechnology*, Vol. 168, No. 4, pp. 303-314, 2013.

344. Lazar, Z., et al., "Hexokinase - A Limiting Factor in Lipid Production from Fructose in Yarrowia Lipolytica", *Metabolic Engineering*, Vol. 26, pp. 89-99, 2014.

345. Sestric, R., et al., "Growth and Neutral Lipid Synthesis by Yarrowia Lipolytica on Various Carbon Substrates Under Nutrient-Sufficient and Nutrient-Limited Conditions", *Bioresource Technology*, Vol. 164, p. 41-46, 2014.

346. Groposila-Constantinescu, D., et al., "Production of Microbial Lipids by Yarrowia Lipolytica", *Romanian Biotechnological Letters*, Vol. 20, No. 6, pp. 10936-10944, 2015.

347. Zhao, C., et al., "Metabolome Analysis and Pathway Abundance Profiling of Yarrowia Lipolytica Cultivated on Different Carbon Sources", *Journal of Biotechnology*, Vol. 206, pp. 42-51, 2015.

348. Bellou, S., et al., "High Lipid Accumulation in Yarrowia Lipolytica Cultivated Under Double Limitation of Nitrogen and Magnesium", *Journal of Biotechnology*, Vol. 234, pp. 116-126, 2016.

349. Kuttiraja, M., et al., "Elucidating the Effect of Glycerol Concentration and C/N Ratio on Lipid Production Using Yarrowia Lipolytica SKY7", *Applied Biochemistry and Biotechnology*, Vol. 180, No. 8, pp. 1586-1600, 2016.

350. Mathiazhakan, K., D. Ayed, and R.D. Tyagi, "Kinetics of Lipid Production at Lab Scale Fermenters by a New Isolate of Yarrowia Lipolytica SKY7", *Bioresource Technol*ogy, Vol. 221, pp. 234-240, 2016.

351. Zhao, L., et al., "Cocoa-Butter-Equivalent Production from Yarrowia lipolytica by Optimization of Fermentation Technology", *American Journal of Biochemistry and Biotechnology*, Vol. 12, No. 4, pp. 196-205, 2016.

352. Bhutada, G., et al., "Sugar Versus Fat: Elimination of Glycogen Storage Improves Lipid Accumulation in Yarrowia Lipolytica", *FEMS Yeast Research*, Vol. 17, No. 3, 2017.

353. Gajdos, P., J.M. Nicaud, and M. Certik, "Glycerol Conversion into a Single Cell Oil by Engineered Yarrowia Lipolytica", *Engineering in Life Sciences*, Vol. 17, No. 3, pp. 325-332, 2017.

354. Sabra, W., et al., "Substrates and Oxygen Dependent Citric Acid Production by Yarrowia Lipolytica: Insights Through Transcriptome and Fluxome Analyses", *Microbial Cell Factory*, Vol. 16, No. 1, pp. 78, 2017.

355. Yuzbasheva, E.Y., et al., "Co-Expression of Glucose-6-Phosphate Dehydrogenase and Acyl-Coa Binding Protein Enhances Lipid Accumulation in The Yeast Yarrowia Lipolytica", *New Biotechnology*, Vol. 39, pp. 18-21, 2017.

356. Yook, S.D., et al., "Efficient Lipid Extraction from the Oleaginous Yeast Yarrowia Lipolytica Using Switchable Solvents", *Renewable Energy*, Vol. 132, pp. 61-67, 2019.

357. Zhang, S., et al., "pH Selectively Regulates Citric Acid and Lipid Production in Yarrowia Lipolytica W29 During Nitrogen-Limited Growth on Glucose", *Journal of Biotechnology*, Vol. 290, pp. 10-15, 2019.

358. Kumar, L.R., et al., "Purified Crude Glycerol by Acid Treatment Allows to Improve Lipid Productivity by Yarrowia Lipolytica SKY7", *Process Biochemistry*, Vol. 96, pp. 165-173, 2020.

359. RStudio, "RStudio: Integrated development environment for R", *RStudio Team*, Boston, 2020.

360. Drévillon, L., M. Koubaa, and E. Vorobiev, "Lipid Extraction from Yarrowia Lipolytica Biomass Using High-Pressure Homogenization", *Biomass and Bioenergy*, Vol. 115, pp. 143-150, 2018.

361. Kolouchova, I., et al., "Production of Palmitoleic and Linoleic Acid in Oleaginous and Nonoleaginous Yeast Biomass", *International Journal of Analyrical Chemistry*, Vol. 2016, pp. 7583684, 2016.

362. Carsanba, E., S. Papanikolaou, and H. Erten, "Production of Oils and Fats by Oleaginous Microorganisms with an Emphasis Given to the Potential of the Nonconventional Yeast Yarrowia Lipolytica", *Critical Reviews in Biotechnology*, Vol. 38, No. 8, pp. 1230-1243, 2018.

363. Magdouli, S., et al., "Sync Between Leucine, Biotin and Citric Acid to Improve Lipid Production by Yarrowia Lipolytica on Crude Glycerol-Based Media", *Biomass and Bioenergy*, Vol. 142, 2020.

364. Chen, J., et al., "Lipid Production from Fed-Batch Fermentation of Crude Glycerol Directed by the Kinetic Study of Batch Fermentations", *Fuel*, Vol. 209, pp. 1-9, 2017.

365. Makri, A., S. Fakas, and G. Aggelis, "Metabolic Activities of Biotechnological Interest in Yarrowia Lipolytica Grown on Glycerol in Repeated Batch Cultures", *Bioresource Technol*ogy, Vol. 101, No. 7, pp. 2351-2358, 2010.

366. Liu, N., K. Qiao, and G. Stephanopoulos, "(13)C Metabolic Flux Analysis of Acetate Conversion to Lipids by Yarrowia Lipolytica", *Metabolic Engineering*, Vol. 38, pp. 86-97, 2016.

367. Dourou, M., et al., "Storage Lipid and Polysaccharide Metabolism in Yarrowia Lipolytica and Umbelopsis Isabellina", *Applied Microbiology and Biotechnology*, Vol. 101, No. 19, pp. 7213-7226, 2017.

368. Bao, W., et al., "Approaches to Improve the Lipid Synthesis of Oleaginous Yeast Yarrowia Lipolytica: A Review", *Renewable and Sustainable Energy Reviews*, Vol. 149, 2021.

369. Gao, R., et al., "Enhanced Lipid Production by Yarrowia Lipolytica Cultured with Synthetic and Waste-Derived High-Content Volatile Fatty Acids Under Alkaline Conditions", *Biotechnology for Biofuels and Bioproducts*, Vol. 13, pp. 3, 2020.

370. Go, A.W., et al., "Biodiesel Production Under Subcritical Solvent Condition Using Subcritical Water Treated Whole Jatropha Curcas Seed Kernels and Possible Use of Hydrolysates to Grow Yarrowia Lipolytica", *Fuel*, Vol. 120, pp. 46-52, 2014.

371. Poli, J.S., et al., "Microbial Lipid Produced by Yarrowia Lipolytica QU21 Using Industrial Waste: A Potential Feedstock for Biodiesel Production", *Bioresource Technol*ogy, Vol. 161, pp. 320-326, 2014.

372. Chang, Y.-H., et al., "A Comparative Study on Batch and Fed-Batch Cultures of Oleaginous Yeast Cryptococcus Sp. in Glucose-Based Media and Corncob Hydrolysate for Microbial Oil Production", *Fuel*, Vol. 105, pp. 711-717, 2013.

373. Ochoa-Estopier, A. and S.E. Guillouet, "D-Stat Culture for Studying the Metabolic Shifts from Oxidative Metabolism to Lipid Accumulation and Citric Acid Production in Yarrowia Lipolytica", *Journal of Biotechnology*, Vol. 170, pp. 35-41, 2014.

374. Robles-Rodriguez, C.E., et al., "Dynamic Metabolic Modeling of Lipid Accumulation and Citric Acid Production by Yarrowia Lipolytica", *Computers & Chemical Engineering*, Vol. 100, pp. 139-152, 2017.

375. Timoumi, A., et al., "Impacts of Environmental Conditions on Product Formation and Morphology of Yarrowia Lipolytica", *Applied Microbiology and Biotechnology*, Vol. 102, No. 9, pp. 3831-3848, 2018.

376. Bellou, S., et al., "Morphological and Metabolic Shifts of Yarrowia Lipolytica Induced by Alteration of the Dissolved Oxygen Concentration in the Growth Environment", *Microbiology*, Vol. 160, No. 4, pp. 807-817, 2014.

377. Papanikolaou, S., et al., "Single Cell Oil Production by Yarrowia Lipolytica Growing on an Industrial Derivative of Animal Fat in Batch Cultures", *Applied Microbiology and Biotechnology*, Vol. 58, No. 3, pp. 308-312, 2002.

378. Zhao, X., et al., "NiMg/Ceria-Zirconia Cylindrical Pellet Catalysts for Tri-reforming of Surrogate Biogas", *Industrial & Engineering Chemistry Research*, Vol. 57, No. 3, pp. 845-855, 2018.

379. Zhao, X., et al., "Tri-Reforming of Surrogate Biogas over Ni/Mg/Ceria–Zirconia/Alumina Pellet Catalysts", *Chemical Engineering Communications*, Vol. 205, No. 8, pp. 1129-1142, 2018.

380. Zhao, X., et al., "Design and Optimization of Nimg/Ceria-Zirconia Catalyst Pellets", *Powder Technology*, Vol. 357, pp. 214-222, 2019.

381. Wachter, P., et al., "Experimental Investigation of Tri-Reforming on a Stationary, Recuperative TCR-Reformer Applied to an Oxy-Fuel Combustion of Natural Gas, Using a Ni-Catalyst", *Energy*, Vol. 212, pp. 118719, 2020.

382. Kantserova, M.R., S.M. Orlyk, and O.D. Vasylyev, "Catalytic Activity and Resistance to Sulfur Poisoning of Nickel-Containing Composites Based on Stabilized Zirconia in Tri-reforming of Methane", *Theoretical and Experimental Chemistry*, Vol. 53, No. 6, pp. 387-394, 2018.

383. Izquierdo, U., et al., "Catalyst Deactivation and Regeneration Processes in Biogas Tri-Reforming Process. The Effect of Hydrogen Sulfide Addition", *Catalysts*, Vol. 8, No. 1, pp. 12, 2018.

384. Anchieta, C.G., E.M. Assaf, and J.M. Assaf, "Effect of Ionic Liquid in Ni/ZrO2 Catalysts Applied to Syngas Production by Methane Tri-Reforming", *International Journal of Hydrogen Energy*, Vol. 44, No. 18, pp. 9316-9327, 2019.

385. Anchieta, C.G., E.M. Assaf, and J.M. Assaf, "Syngas Production by Methane Tri-Reforming: Effect of Ni/Ceo2 Synthesis Method on Oxygen Vacancies and Coke Formation". *Journal of CO2 Utilization*, Vol. 56, pp. 101853, 2022.

386. Yabe, T., et al., "Ni-Mg Supported Catalysts on Low-Temperature Electro-Catalytic Tri-Reforming of Methane with Suppressed Oxidation", *ACS Catalysis*, Vol. 8, No. 12, pp. 11470-11477, 2018.

387. Fedorova, Z.A., M.M. Danilova, and V.I. Zaikovskii, "Porous Nickel-Based Catalysts for Tri-Reforming of Methane to Synthesis Gas: Catalytic Activity", *Materials Letters*, Vol. 261, pp. 127087, 2020.

388. García-Vargas, J.M., et al., "Precursor Influence and Catalytic Behaviour of Ni/CeO2 and Ni/SiC Catalysts for the Tri-Reforming Process", *Applied Catalysis A: General*, Vol. 431-432, pp. 49-56, 2012.

389. Soares, A.N.D.B., et al., "The Effect of Carbon Dioxide in the Feed Stream of Tri-Reforming of Methane Process Compared to the Partial Oxidation of Methane", *Environmental Science and Pollution Research*, Vol. 30, No. 7, pp. 19111-19119, 2022.

390. De C. Roseno, K.T., et al., "Tri-Reforming of Methane over NdM0.25Ni0.75O3 (M = Cr, Fe) Catalysts and the Effect of CO2 Composition", *Catalysis Letters*, Vol. 151, No. 12, pp. 3639-3655, 2021.

391. Jiang, H., et al., "Preparation of Ni/MgXTi1−XO Catalysts and Investigation on Their Stability in Tri-Reforming of Methane", *Fuel Processing Technology*, Vol. 88, No. 10, pp. 988-995, 2007.

392. García-Vargas, J.M., et al., "Catalytic and Kinetic Analysis of the Methane Tri-Reforming over a Ni–Mg/β-SiC Catalyst", *International Journal of Hydrogen Energy*, Vol. 40, No. 28, pp. 8677-8687, 2015.

393. Solov'ev, S.A., et al., "Tri-Reforming of Methane on Structured Ni-Containing Catalysts", *Theoretical and Experimental Chemistry*, Vol. 48, No. 3, pp. 199-205, 2012.

394. Lino, A.V.P., E.M. Assaf, and J.M. Assaf, "Adjusting Process Variables in Methane Tri-reforming to Achieve Suitable Syngas Quality and Low Coke Deposition", *Energy & Fuels*, Vol. 34, No. 12, pp. 16522-16531, 2020.

395. Sun, D., et al., "Effect of O2 and H2O on the Tri-Reforming of the Simulated Biogas to Syngas over Ni-Based SBA-15 Catalysts", *Journal of Natural Gas Chemistry*, Vol. 19, No. 4, pp. 369-374, 2010.

396. Vita, A., et al., "Biogas as Renewable Raw Material for Syngas Production by Tri-Reforming Process over NiCeO2 Catalysts: Optimal Operative Condition and Effect of Nickel Content", *Fuel Processing Technology*, Vol. 127, pp. 47-58, 2014.

397. Majewski, A.J. and J. Wood, "Tri-Reforming of Methane over Ni@SiO 2 Catalyst", *International Journal of Hydrogen Energy*, Vol. 39, No. 24, pp. 12578-12585, 2014.

398. Pino, L., et al., "Hydrogen Production by Methane Tri-Reforming Process over Ni–Ceria Catalysts: Effect of La-Doping", *Applied Catalysis B: Environmental*, Vol. 104, No. 1-2, pp. 64-73, 2011.

399. Kumar, K.D.P.L., et al., "Enhanced CO2 Utilization via Methane Tri-Reforming over Ru Incorporated Co/MgO-Al2O3 Vatalyst: Influence of La and Ce Promoters", *Journal of Environmental Chemical Engineering*, Vol. 9, No. 5, pp. 105949, 2021.

400. Singha, R.K., et al., "Energy Efficient Methane Tri-Reforming for Synthesis Gas Production over Highly Coke Resistant Nanocrystalline Ni–ZrO2 Catalyst", *Applied Energy*, Vol. 178, pp. 110-125, 2016.

401. Da Costa, K.Ś., et al., "Understanding of Tri-Reforming of Methane over Ni/Mg/Al Hydrotalcite-Derived Catalyst for CO2 Utilization from Flue Gases from Natural Gas-Fired Power Plants", *Journal of CO2 Utilization*, Vol. 42, pp. 101317, 2020.

402. Sun, L.-z., et al., "Tri-Reforming of Coal Bed Methane to Syngas over the Ni-Mg-ZrO2 Catalyst", *Journal of Fuel Chemistry and Technology*, Vol. 40, No. 7, pp. 831-837, 2012.

403. Yoo, J., et al., "Hydrogen Production by Tri-Reforming of Methane over Nickel–Alumina Aerogel Catalyst", *Journal of Molecular Catalysis A: Chemical*, Vol. 410, pp. 74-80, 2015.

404. Kumar, R., et al., "Effect of Support Materials on the Performance of Ni-Based Catalysts in Tri-Reforming of Methane", *Fuel Processing Technology*, Vol. 186, pp. 40-52, 2019.

405. Walker, D.M., et al., "Synthesis Gas Production to Desired Hydrogen to Carbon Monoxide Ratios by Tri-Reforming of Methane Using Ni–MgO–(Ce,Zr)O2 Catalysts", *Applied Catalysis A: General*, Vol. 445-446, pp. 61-68, 2012.

406. Pandey, A., et al., "Tri-reforming of CH_4 over a Thermally Stable and Carbon-Resistant Nanonickel Metal Catalyst Dispersed on Mesoporous-Zirconia", *Industrial & Engineering Chemistry Research*, Vol. 63, No. 2, pp. 1000-1012, 2024.

407. Nascimento, J.P., et al., "Investigation of the Deactivation Behavior of MeMo/La_2O_3-Al_2O_{3-} and MeMo/Nb_2O_5 Supported Catalysts (Me = Pt, Ni, and Co) in Tri-Reforming of Methane", *Energy & Fuels*, Vol. 37, No. 5, pp. 3836-3853, 2023.

408. Erdem Günay, M. and R. Yıldırım, "Recent Advances in Knowledge Discovery for Heterogeneous Catalysis Using Machine Learning", *Catalysis Reviews*, Vol. 63, No. 1, pp. 120-164, 2021.

409. Wang, S., G.Q. Lu, and G.J. Millar, "Carbon Dioxide Reforming of Methane to Produce Synthesis Gas over Metal-Supported Catalysts: State of the Art", *Energy & Fuels*, Vol. 10, No. 4, pp. 896-904, 1996.

410. Li, B. and S. Rangarajan, "A Conceptual Study of Transfer Learning with Linear Models for Data-Driven Property Prediction", *Computers & Chemical Engineering*, Vol. 157, pp. 107599, 2022.

411. Wen, M., et al., "Improving Machine Learning Performance on Small Chemical Reaction Data with Unsupervised Contrastive Pretraining", *Chemical Science*, Vol. 13, No. 5, pp. 1446-1458, 2022.

412. Weiss, K., T.M. Khoshgoftaar, and D. Wang, "A Survey of Transfer Learning", *Journal of Big Data*, Vol. 3, No. 1, pp. 9, 2016.

413. Ni, J., et al., "An Improved Transfer Learning Model for Cyanobacterial Bloom Concentration Prediction", *Water*, Vol. 14, No. 8, pp. 1300, 2022.

414. Al Ibrahim, E. and A. Farooq, "Transfer Learning Approach to Multitarget Temperature-Dependent Reaction Rate Prediction", *The Journal of Physical Chemistry A*, Vol. 126, No. 28, pp. 4617-4629, 2022.