Optimization Based Framework for The Predication of Biomass Liquefaction using ML Methods Integrated with Metaheuristic Techniques

By

Attiq ur Rehman

(Registration No: 00000362453)

Department of Process System Engineering

School of Chemical and Materials Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Optimization Based Framework for the Predication of Biomass Liquefaction using ML Methods Integrated with Metaheuristic Techniques

By

Attiq ur Rehman

(Registration No: 00000362453)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Process System Engineering

Supervisor: Dr. Muhammad Nauman Aslam Khan

School of Chemical and Materials Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by Mr **Attiq Ur Rehman** (Registration No 00000362453), of School of Chemical & Materials Engineering (SCME) has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.
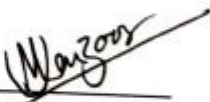
Signature: _____

Name of Supervisor: **Dr M. Nouman Aslam Khan**

Date: _____ July 01, 2024 _____

Signature (HOD): _____

Date: _____ 1/7/24 _____

Signature (Dean/Principal): _____

Date: _____ 1/7/24 _____

TH - 1

$P3E-04-2021$

National University of Sciences & Technology (NUST)

MASTER'S THESIS WORK

Formulation of Guidance and Examination Committee (GEC)

ame: **Attiq Ur Rehman**   NUST Reg No: 00000362453

epartment: **Department of Chemical Engineering**   Specialization: Master of Science in Process System Engineering

redit Hour Completed: 21.0   CGPA: 3.0

Course Work Completed

S/No:	Code:	Title:	Core/Elective:	CH:	Grade:
1.	PSE-801	Process Systems Theory	Compulsory	3.0	B
2.	PSE-852	Process Modelling and Simulation	Compulsory	3.0	B
3.	TEE-820	Process Intensification	Elective	3.0	B
4.	RM-898	Research Methodology	Additional	2.0	Q
5.	PSE-823	Advanced Process Dynamics and Control	Compulsory	3.0	B
6.	PSE-802	Optimization and Decision Analysis	Compulsory	3.0	B
7.	CSE-801	Computational Fluid Dynamics	Elective	3.0	B
8.	CHE-814	Product Technology	Elective	3.0	B

ate 19 - Oct - 2022   Student's Signature

Thesis Committee

Name: Muhammad Nouman Aslam Khan (**Supervisor**)   Signature

Department: Department of Chemical Engineering

Name: Salman Raza Naqvi (Internal)   Signature

Department: Department of Chemical Engineering

Name: Dr. Nouman Ahmad (Internal)   Signature

Department: Department of Chemical Engineering

ate: 19 - Oct - 2022   Signature of Head of Department:

APPROVAL

ate: 19 - Oct - 2022   Signature of Dean/Principal:

School of Chemical & Materials Engineering (SCME) (SCME) H-12 Campus.

TH – 4

Form: TH-04

National University of Sciences & Technology (NUST)

MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by

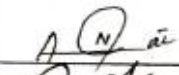Regn No & Name: **00000362453 Attiq Ur Rehman**

Title: **Optimization Based Framework for the Predication of Biomass Liquefaction using Machine Learning Methods Integrated with Metaheuristic Techniques.**

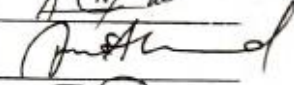Presented on: **27 Jun 2024**   at: **1500 hrs in SCME (Seminar Hall)**

Be accepted in partial fulfillment of the requirements for the award of Master of Science degree in **Process Systems Engineering**.

Guidance & Examination Committee Members

Name: **Dr Muhammad Adnan**   Signature: _____

Name: **Dr Nouman Ahmad**   Signature: _____

Supervisor's Name: **Dr M. Nouman Aslam Khan**   Signature: _____

Dated: **27/06/2024**

Head of Department
Date 28/6/24

Dean/Principal
Date _____

School of Chemical & Materials Engineering (SCME)

AUTHOR'S DECLARATION

I <u>Attiq ur Rehman</u> hereby state that my MS thesis titled "<u>Optimization Based framework for the predication of Biomass liquefaction using Machine Learning Methods Integrated with Metaheuristic Techniques</u>" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: <u>Attiq ur Rehman</u>

Date: <u>19 June 2024</u>

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "Optimization Based framework for the predication of Biomass liquefaction using Machine Learning Methods Integrated with Metaheuristic Techniques" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: _____

Name: Attiq ur Rehman

DEDICATION

" To my very Supportive, Loving, and Caring Family."

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

-

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ML	Machine Learning Methods
DT	Decision Tree
SVM	Support Vector Machines
GPR	Gaussian Process Regression
EL	Ensembled Learning
ET	Ensembled Tree
GA	Genetic Algorithm
R2	Coefficient of determination PSO Particle Swarm Optimization
RMSE	Root mean square error
FC	Fixed carbon
MC	Moisture Content
VM	Volatile matter

ABSTRACT

The liquefaction of microalgae yields bio-oil, that is a potential alternative fuel. However, it is a complicated and challenging process to use the method of experimentation to investigate the relationship between liquefaction conditions of, ultimate, and proximal analysis with bio-oil production. so, an efficient and well-organized model must be created to reliably predict the effect of input parameters on the bio-oil yield. This study utilizes a novel PSO-based and genetic algorithm-based approach for feature selection and hyperparameter optimization. Four different machine learning models were constructed and compared using these optimization methods. It was found that Ensembled Tree model performed better and R2=0.6801, RMSE=1.7226 for GA and the values of R2 (Coefficient of determination) = 0.983 and RMSE (Root mean Square Error) = 1.4548 while using PSO performed better and highly recommend based features selection. The result of SVM was the worst one R2 = 0.7483, RMSE =10.0063 for PSO and R2 = 0.6068, RMSE = 35.3462 for GA. For the PSO-based study, the R2 values for DT and GPR were 0.8125 and 0.9309, respectively, while the R2 values for the GA-based methods were 0.8757 and 0.9270.

Keywords: Oak Tree, Liquefaction, ML, Genetic Algorithm, Particle Swarm Optimization

GRAPHICAL ABSTRACT
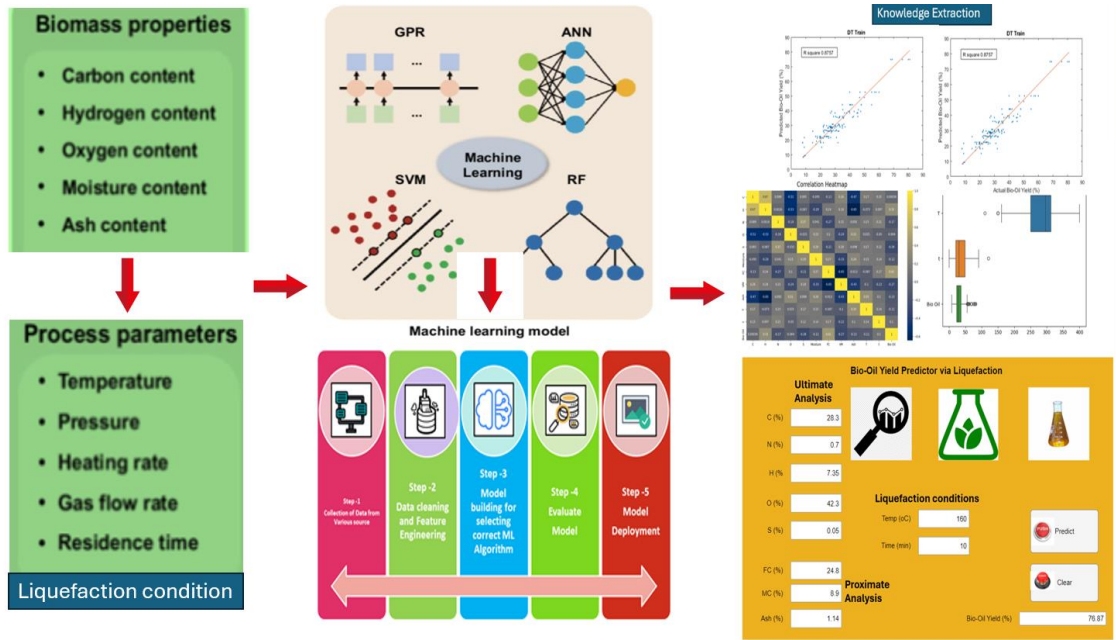
CHAPTER 1:   INTRODUCTION

1.1   Background

Energy viewpoint - all have advantages and disadvantages, and they did not create the massive energy storage required to dispatch the intermittent energy provided by wind and solar photovoltaic.[1]

Energy is essential to the economic prosperity of a country and its society. It is currently the focus of great interest in the political, economic, military, and other areas across the world. [2]

A mismatch between global energy supply and demand has emerged in recent years as the global economy has grown and modernised. Climate and resource limits have also worsened, posing severe challenges to global energy supplies. To achieve carbon neutrality, the global energy system has become more diverse, and the amount of renewable energy will continue to rise as electrification develops.[3]

Developing nations tend to attract highly polluting corporations due to the opportunities given by their growth objectives, rendering them particularly vulnerable to air pollution. Such a scenario is mostly caused by inadequate environmental regulations and affordable or plentiful labour forces in emerging nations compared to wealthy nations. As a result of established nations' strict environmental rules, emerging countries become pollution hotspots. Furthermore, poor air quality increases the global burden of respiratory ailments, putting additional financial demand on governments and citizens alike to care for their health. Improved air quality will alleviate suffering while saving money on both public and private health care.[4]

The use of biomass as a fuel for energy production has raised controversy about the associated greenhouse gas emissions. Proponents of the fuel believe that these emissions are neutral in terms of CO_2 release, however opponents note that CO_2 emissions are emissions regardless of source, and that utilising biomass for energy production may result in misuse of forest resources.[5]

The carbon neutrality of burning biomass for energy does not imply that there is no CO_2 emissions into the atmosphere. This carbon neutrality is closely tied to the carbon cycle itself. In other words, it implies that the carbon released into the atmosphere is precisely the same as what was taken by the plant species over its lifespan, and that it always corresponds to a lifetime in which the atmosphere matches the present.[6]

On the contrary, the use of fossil fuels encourages the emission of carbon that was collected and stored by plant species, but during a time in the planet's geological history when the atmosphere also contained CO_2 concentrations that were far greater than those now.[7]

While using biomass as fuel may be a viable backup plan, several limits must be considered in order to make the greatest use of this resource feasible. The purpose of this study is to look at the possibility of using biomass as an energy source, with a specific emphasis on the use of residual woody biomass from agriculture and forestry management activities like pruning. Thus, the research is conducted in two steps: First, a PEST analysis is conducted to see if energy can be recovered from woody wastes in agroforestry. Second, based on the findings, any negative or weaker characteristics that were detected are explored using a SWOT analysis to offer a model of resource exploitation,

By 2050, the global population is expected to reach 9.9 billion, resulting in increased urbanisation and economic levels, as well as a demand for cattle products. The need for animal feed will nearly double globally.[9]

Removing the usage of animal products from the feed production process can have detrimental effects on the environment, the economy, and society. Some of these effects include pollution of the natural ecosystem, waste of resources from slaughterhouses, and increased feed production costs as a result of the demand for higher value-added components.[10] Hydrothermal liquefaction (HTL) is one of these methods that can effectively turn wet biomass into energy-density biocrude and its byproducts (aqueous phase, gas phase, and hydro-char) without the need for a pre-drying procedure.[11] It is also acknowledged that a number of sophisticated HTL pathways have been created to valorise manure wastes, and in recent years, the reactors that correspond to these

pathways have also been studied. Fast-HTL is a revolutionary HTL technology that can save energy and increase biocrude production by rapidly heating materials at temperatures above 400 °C. [12]

HTL is seen as a promising technology for creating bio-crude, which may then be processed into advanced biofuels and value-added products, from biomass and organic wastes, particularly wet organic wastes. The thermochemical process known as HTL requires high temperatures (250-375 °C), pressures larger than water's saturation pressure, and a rapid reaction time. Given the difficult working conditions, HTL has been shown as a micropollutant decomposition technique.[13]

Most research focused on the conversion of specific food waste kinds, such as fruits, cheese, and meats. This study investigates a real food waste combination as an HTL feedstock. Nowadays, several facilities co-digest FW and SS to generate biogas. This study also investigates the co-HTL of FW with main and secondary sludge in the WWTP-handled FW scenario. FOG, food waste, primary and secondary sludges, and digestate all have distinct biochemical makeup. Thus, this research intends to achieve numerous objectives in order to completely comprehend the potential of the hydrothermal liquefaction (HTL) process in wastewater treatment plants (WWTPs). These include determining if synergies exist in different sludge mixtures, as shown in other feedstocks, providing an overview of the bio-crude potential of various sludge kinds and WWTPs, and building simpler models to help with bio-crude production forecast. Furthermore, this study seeks to investigate the inorganic composition of hydrochar for possible nutrient recovery, as well as the impacts of various sludge types and WWTP on the HTL aqueous phase.[14]

Calculating H2 yield output and its connection to input parameters (proximate, final composition, and gasification conditions) using traditional methods requires a significant amount of work, expense, and time. To effectively study the combined effects of supercritical water gasification parameters on H2, artificial intelligence techniques such as deep learning (DL), machine learning (ML), and data mining must be used to investigate their behaviour while taking SS composition into account. Machine learning-

3

based models give a feasible solution by using process data to generate a model for automation, monitoring, and optimisation.[15]

Machine learning techniques assess patterns in observations of the same classification and identify features that differentiate the observations of different groups.

ML techniques have proven to be reliable alternatives to the traditional modelling techniques for studying and comprehending complex conversion processes. They can model complex problems by leveraging a series of interconnections between the input and output parameters. Recently, the application of ML approaches to model thermochemical biofuel conversion processes have increased significantly.

The applications include predicting product production and composition, as well as kinetic parameters and caloric values. The ML models created were also utilised to estimate the best circumstances for intended production. The preprocessing or normalisation of input variables with various magnitudes is a hyperparameter that helps to improve the performance of machine learning models. However, several did not mention the strategies employed to prevent overfitting, which is a critical hurdle against any model improvement at all..[16,17]

1.2 Animal and Plant Biomasses

China's livestock business is rapidly expanding, leading to the yearly slaughter of 7×108 pigs. Animal fatalities, whether intentional or inadvertent, will undoubtedly occur, resulting in a large number of corpses that may be damaging to both human health and the environment. The amount of dead pigs generated annually is staggering. Before being marketable in 2017, over 60 million pigs died due to illness, natural causes, or other reasons.[18]

Algal biomass is divided into two main classes. (1) There are two types of algae: (1) microalgae, which are invisible to the human eye, and (2) macroalgae, which are visible. Algae can supplement lignocellulosic biomass as a feedstock for biofuels. Lipid production by algae has long been the subject of research. On the other hand, high lipid algae require careful cultivation as they grow over time. The processing of algal biomass according to its lipid content is depicted in Figure 1. Algal blooms, which happen frequently in many bodies of water, can provide algal biomass. Algal blooms usually have negative effects on the local economy and ecology. Estimates indicate that eutrophication of river systems and marine algal blooms cause yearly.
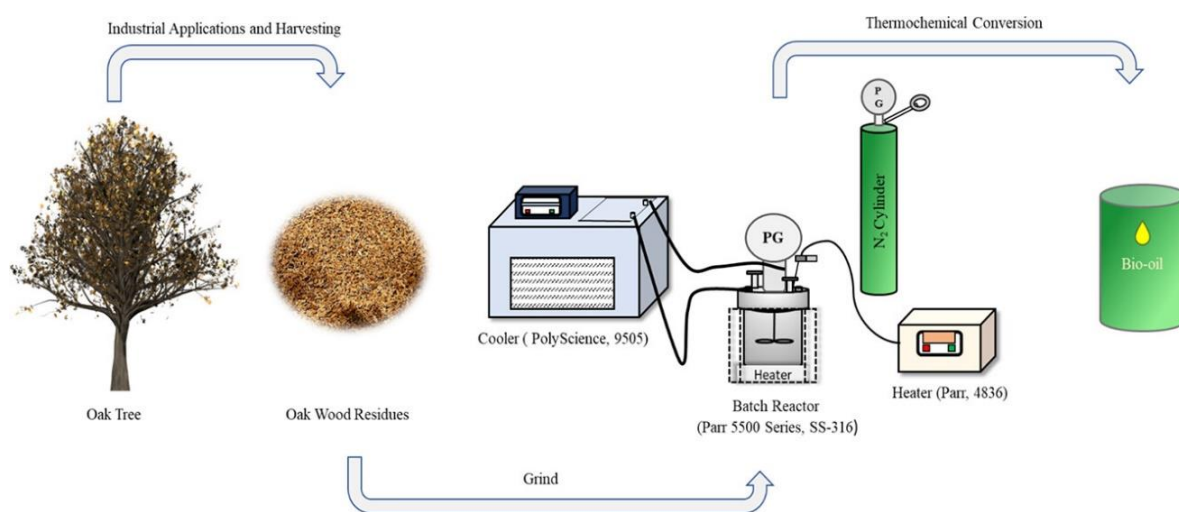
Figure 1.1: *Processing of Oak Tree biomass into bio-oil [18]*

economic losses in the US of close to $2.2 billion and $100 million, respectively.[18]

Figure 1.2: *Different type of Algae [18]*

According to the pigments that give them colour, Figure 2 displays pictures of various microalgae species. While being mostly employed in culinary products, macroalgae are a wonderful source of biomedical and pharmacological elements such as fucoidan, fucoxanthin, laminarin, mannitol, and Phyto colloids. Conversely, microalgae are single-celled microorganisms. It has 2000 species in the world that includes cyanobacteria, Chlorophyceae, Xanthophyte, Charophyceae, and Bacillariophyceae and many others. Because of their antibacterial, anti-inflammatory, and biotechnological potential, microalgae have so far been extensively used in the production of products including fertilizers, pharmaceutical medicine, chemicals, food products, and biofuels. The yearly creation of microalgae continuously expanded from 500 to 20,000 tons between 1984 and 2016, indicating that microalgae are consistently used as feedstock, raw materials, or products. Unlike macroalgae, the produced microalgal feedstock from harvesting and cultivation is typically disseminated in suspension, necessitating surface dewatering and harvesting based on its size, density, cell charge, and desired yield. Figure 3 shows the powdered form of microalgae.

Figure 1.3: *Powdered Microalgae [18]*

CHAPTER 2: LITERATURE REVIEW

The demand for a sufficient standard of living is driven by the expanding human population and developing economies, which in turn causes the depletion of fossil fuels deposits, increased emissions of green house gas and other undesirable gases, and rising energy costs. Renewable energies sources can thus replace fossil fuels and are also clean energy sources because fossil fuel reserves are essentially endless and have been steady since the beginning of time [21]. Geothermal, wind, and hydroelectric energy are examples of renewable energy sources. Other than that, biomass is an important source of clean energy [22]. Algal biomass, agricultural residue, non-woody and woody plants, municipal solid waste, and forest residue are examples of biomass resources. Additional categories for solid waste from municipalities include plastics, woods, papers, cloth, rubbers, ashes, metal, diaper, and more materials [23].

Algal biomasses can be divided into two main branches: macroalgae, which are somewhat larger than farmers, and microalgae, which are invisible to the human eye [24]. A further classification of microalgae yields 3000 recognised species. Less than a hundred of them have been thoroughly examined and identified. Spirulina, Chlorella, Nanochloropsis, Dunaliella, and a few more species are examples of microalgal species [25]. Being photosynthetic organisms, they may grow and reproduce in a variety of settings and don't require fertile soil or clean water to be farmed [26]. Their composition allows for a wide range of uses in the production of high-value goods, such as chemicals, cosmetics, and human nutrition. In addition, microalgae can be utilised for biofuel [27]. Microalgae are used because of their high lipid content, which makes them a viable option for the manufacture of bio-oil [28]. Bio-oil can be produced from microalgae using a variety of thermochemical conversion processes, including gasification, pyrolysis, liquefaction, combustion, and torrefaction [20]. The list of thermochemical methods for treating algae is displayed in Figure 4.

Pyrolysis is the process of using heat to break down biomass at a high temperature in an inert atmosphere. The main outcomes of pyrolysis's are biochar, bio-oil, and syngas.

Power generation, mechanical activities, and other chemical industries can use pyrolysis bio-oil [29]. Pyrolysis can be categorised as quick or slow based on many operational criteria such as heating rate, residence time, and others [30]. Various researchers have studied the effects of variables employed in microalgae pyrolysis on laboratory-produced bio-oil. Yang et al. [31] discuss the impact of feed type and pyrolysis temperature on the volume of bio-oil generated during the pyrolysis of several microalgae biomasses. In their investigation of the effects of pyrolysis duration and temperature (T) on bio-oil yield, Kim et al. [32] studied the pyrolysis of Daniella microalgae at various temperatures and residence times. In their discussion of the effects of temperature and N_2 flow rate on the pyrolysis of microalgae, Muradov et al. [33] discovered that the maximum yield (40.6%) was achieved at a pyrolysis temperature of 500oC and a N_2 flowrate of 80 ml/min. Many researchers have worked on applying artificial intelligence in the field of biomass on different thermochemical conversion techniques. The number of articles on machine learning in the fields of gasification and pyrolysis is displayed in Figure 5. Machine learning and optimisation were used by Zahid et al. [35] to forecast the production of biofuel from the pyrolysis of biomass. For feature selection and hyperparameter optimisation, they employed genetic algorithms. For additional prediction, four distinct machine learning models—Random Forest, Decision Tree, Support Vector Machine, and MLR—were used. It was discovered that the random forest's coefficient of determination was 0.96.
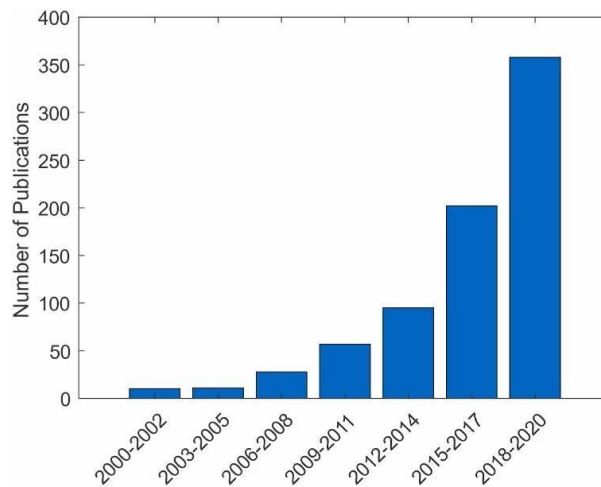
Figure 2.1: *Use of machine learning in modelling applications for liquefaction [36]*
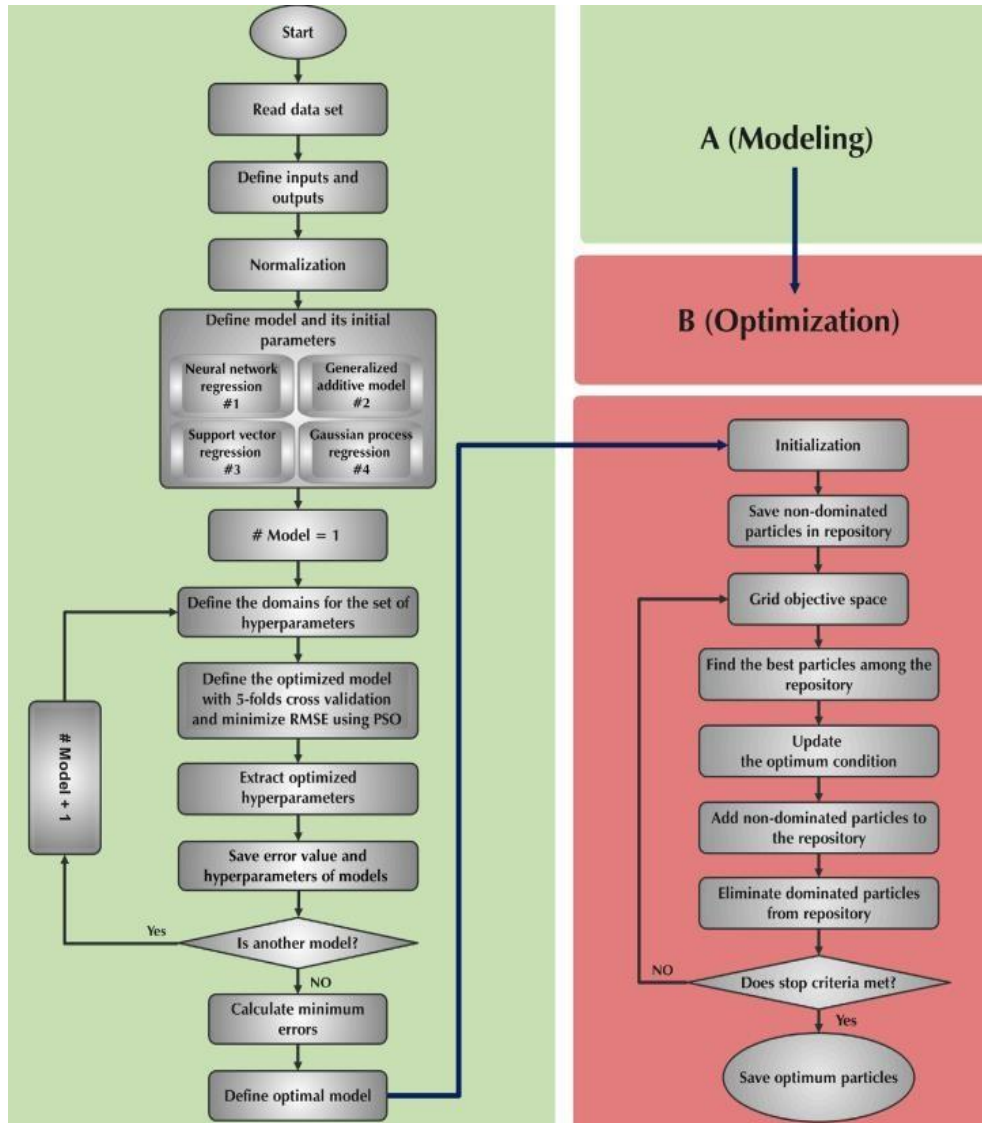
Figure 2.2: *Machine Learning Models development [36]*

Similarly, Shafizadeh et al. [36] investigated hydrothermal liquification of biomass using Machine Learning. They used four distinct ML models: neural network regression, Gaussian process regression, support vector regression, and generalised additive model. The hyperparameters of the machine learning models used were adjusted using a novel approach known as particle swarm optimisation. The optimal response parameters and biomass composition were discovered by combining multi-objective particle swarm optimisation (MOPSO), a multi-objective evolutionary technique, with the top-performing machine learning model's goal function. Finally, without requiring any machine learning

knowledge, a simple script was developed based on the chosen machine learning model, allowing for the avoidance of unnecessary expensive and time-consuming testing. The flowchart for the development of machine learning in this study is displayed in Figure 6.

To forecast the hydrogen levels and bio-oil from biomass, Tang et al. [37] used an ML model. They employed Multiple Linear Regression and Random Forest, two distinct machine learning models. Compared to MLR, the RF method's prediction performance showed a higher potential for generalisation. Furthermore, the partial dependency and feature importance investigated with the RF technique have revealed the fundamental elements as well as the modes that influence the properties of oil. Additionally, Zhang et al. [38] employed machine learning models designed to forecast the properties of biofuels derived from biomass pyrolysis. The pyrolytic biofuel was evaluated in detail, with forecasts for its calorific value, viscosity, components, and amount (yield).

Meanwhile, it was established that machine learning could forecast the viscosity, H/C, and calorific value of biofuel. Using machine learning methods, the yield, calorific value, viscosity, O/C, and H/C of bio-oil were predicted in accordance with the feedstock compositions and pyrolysis factors, while a parallel study was conducted. This survey provides practical suggestions for forecasting biomass-produced biofuel properties and composition under different pyrolysis circumstances. In addition, Zhang et al. [39] created machine learning for the manufacture of biofuels using hydrothermal liquification utilising algae as a feedstock. To estimating Yield oil, O_oil, and N_oil, single and multiple targets are used. Models called GBR and

Additionally, two optimization policies (forward and reverse optimizations) were applied for presenting the perfect feedstock constituents and respective HTL condition, as well as adapting HTL states for certain alga (e-g Chlorella in this study), to help with the desired bio-oil production. Figure 7 shows the results of the Features importance between the operating conditions and Algae composition in this study.
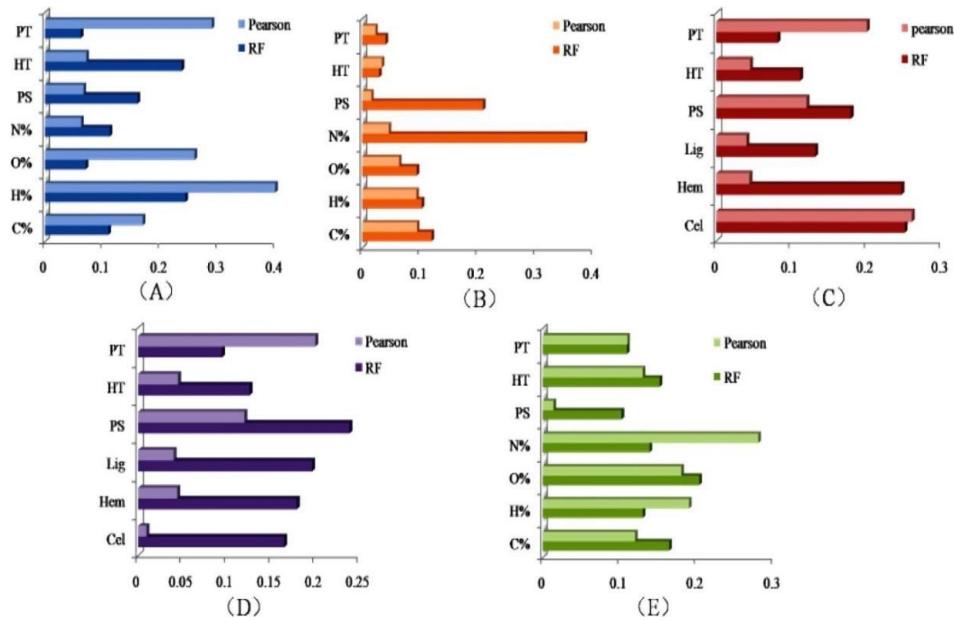
Figure 2.3: *Features Importance of Input Parameters [39]*

Yang et al. [27] [37] used machine learning to predict the oxygen content and biofuel yield of biomass based on pyrolysis parameters and composition. This study investigates the effects of biomass feedstocks and pyrolysis parameters on the pyrolysis process. The datasets from published articles included a significant number of biomass pyrolysis test results. Six RF algorithm-based models were created and trained after the characteristics were decreased to determine their forecast accuracy for biofuel production and oxygen concentration.

Additionally, the trained prediction model performed an significant evaluation of the key features, and the partial dependence analysis (PDA) was utilized to assess the influence of the input features on the target variables. Figure 8 illustrates the methodology followed within this research. Based on the regression coefficient (R2) for the testing data, training predictive performances are evaluated According to the findings, the Proximate-Yield model (R2 1/4 0.925) performs best at forecasting yield of biofuel, while the Ultimate-O model (R2 1/4 0.895) performs best at predicting the oxygen content of bio-oil. The interior data of the biomass is very significant than the pyrolysis factors for predicting the content of oxygen of biofuel, while the heating rate dominates the most importance for predicting biofuel output, according to the analysis of feature importance.

12
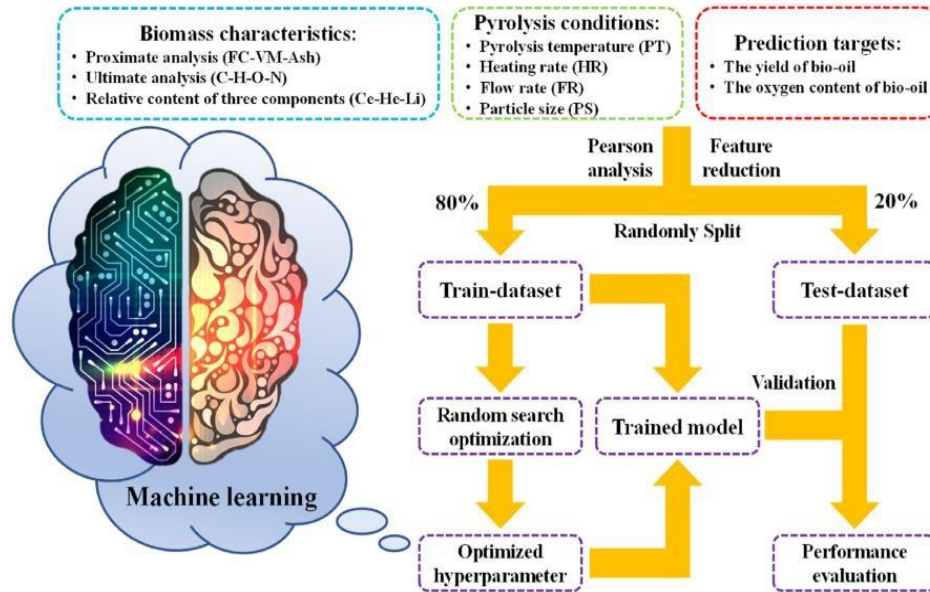
Figure 2.4: *Methodology [37]*

Nonetheless, the customary methods entail costly, costly, and lengthy extensive experimental work to figure out the amount of bio-oil and its connection to the microalgae's composition and pyrolysis circumstances. Consequently, in order to accurately estimate the combined influence on the bio-oil output.

Machine learning should be used to analyse microalgae pyrolysis, taking into account the combined effects of pyrolysis settings and feedstock composition [35]. As artificial intelligence progressed, many unique and cutting-edge strategies that handle both linear and non-linear issues were included into traditional processes [40]. Machine learning techniques such as SVM, DT, MLR, ANN, and RF have previously been effectively used to biomass pyrolysis and gasification due to their higher prediction potential than traditional approaches. [41–42].

2.1   Objectives

The important objectives of this study are given:

➢ Data collection from relevant literature.

➢ Development of Machine Learning models

➢ GA &PSO for features selection.

➢ PSO and GA for the optimization ML models hyperparameters.

➢ Finding the optimum values of input parameters

2.2 Research Justification and Relevance to National

Potential energy policy and climate change are important motivators for expanded biomass consumption worldwide, particularly in developing nations such as Pakistan. One of the main reasons for Pakistan's energy issue is the rising cost of fossil fuels, which are used by local energy firms and household consumers to generate power and heat using boiler fuel and natural gas. Bioenergy has the capacity to produce more energy in addition to taking the place of petroleum-based products with renewable ones. Because fossil fuels are running out and there are environmental issues involved, biofuel made from carbon-neutral biomass feedstocks can be a good substitute.

Various experimental studies have been conducted on the area of biomass pyrolysis produced from Algal feedstock. However, the use of emerging techniques such as Machine Learning can not only save time and resources but can also meet with the requirements of industrial revolution 4.0.

2.3 Thesis Outline

This thesis is structure as follows, in chapter 1 the discussion of introduction of the research study, followed by the literature review. In chapter 3, Overview of Machine learning models along with Optimization algorithms is presented, which is followed by results of the study in chapter 4. Chapter 5 concluded the thesis.

CHAPTER 3: OVERVIEW OF MACHINE LEARNING MODELS & OPTIMIZATION ALGORITHMS

3.1 Machine Learning models

This study generated five different machine learning models. This study used artificial neural networks, decision trees, ensembled learning trees, support vector machines, and Gaussian process regression to develop a model for forecasting the production of bio-oil from biomass by liquefaction. Using MATLAB software and machine learning models.

3.2 Decision Tree (DT)

DT is one of the most used and reliable machine learning models. It depicts solutions and their corresponding results as a tree. Decision trees have several nodes, with the root node being the highest or first. Internal nodes, which contain just one child, are used to represent attribute testing. Based on test findings, the classification algorithms branch and test until they reach the leaf node.[43].

Decision trees have several advantages, including ease of use and interpretation, the ability to determine the optimal value to fill in data gaps, suitability for both regression and classification problems, and high efficiency due to the tree traversal method's performance. Furthermore, over-fitting concerns based on the ensemble technique may develop, and Random Forest is the solution. The key disadvantages of DT include its volatility, difficulties maintaining tree size, and inclination to develop locally optimum rather than globally optimal solutions.[44].

3.3 Basic Concepts

Figure 9 displays a pure DT model with a single binary target variable Y (0 or 1) and two successive variables (x1, x2) that range from 0 to 1. The primary components of a decision tree model are branches and nodes, and the most important phases in model development are halting, splitting, and pruning.[45].
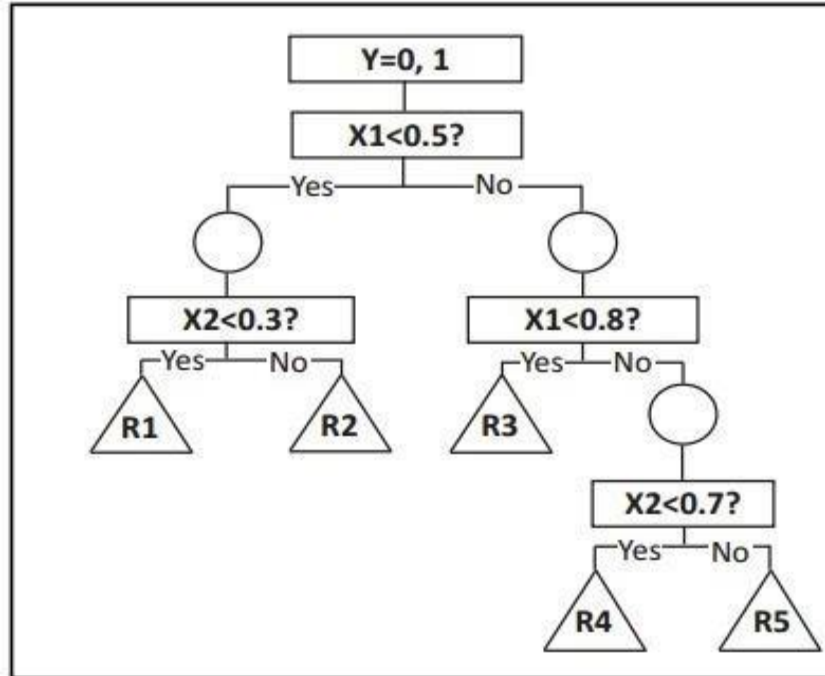
Figure 3.1: *Structure of Decision Tree [45]*

3.4   Nodes

Three categories are used to organise nodes. (1) A choice that will influence the division of a data set into two or more jointly restricted groups is represented by the root node, also known as the decision node. Internal nodes, which are also called chance nodes, represent one of the available alternatives at that particular place in the tree assembly.

The topmost edge of an internal node is connected to its corresponding parent node, while the bottommost edge of an internal node is connected to leaf nodes. (3) Leaf nodes, sometimes referred to as end nodes, characterise the outcome of a sequence of decisions or events.

3.5   Branches

Branches are random events or occurrences that occur from root and internal node. A decision tree model is erected using a branch pyramid. All tracks from the root to the leaf do not symbolize a law of decision classification. These DT routes are often known as 'if- then' rules.

3.6   Advantages

Decision tree has the following advantages when used in research findings.

➢ Divides original input variables into substantial subgroups to simplify complex interactions between target variables and input variables.

➢ Simple to comprehend and interpret.

➢ A non-parametric strategy that creates no statements about the distribution.

➢ It is simple to manage incomplete data without resorting to imputation.

The decision tree approach, like any analytical tools, has limits that user should be aware of. The key downside is that it is susceptible to overfitting and underfitting, especially when used with a limited data set. This issue has the potential to impair the generalizability and resilience of the resulting models. Another potential issue is that substantial correlation between potential input factors may result in the selection of variables that increase statistics of the model but are not necessarily related to the desired outcome. As a result, while evaluating DT models and using the outcomes of these models to establish causal hypotheses, one must use caution.

3.7   Support Vector Machine (SVM)

SVM is a well-liked machine learning model that may be applied to a variety of problems, including regression and classification. In an infinite or high-dimensional space, SVM generates a single hyperplane or a collection of them. Given that the classifier's error diminishes with increasing margin, the hyper-plane farthest from the training data points has a high separation [46]. The length difference between classes is known as the margin. SVM performs incredibly well when used to linear or non-linear dataset problems. The support vector machine technique uses kernel types such polynomials, linear radial base functions, and sigmoid for a model prediction. The Kernel function is used to calculate the degree of similarity between two data points. This means that the kernel type hyperparameter needs to be changed. SVM performs poorly when data have noises, like overlapping target classes [47]. The application of an SVM to a categorization problem is

divided into two stages: training and forecasting. In training, the SVM is fed a dataset with every single sample being a static-length vector. Additionally, each sample must be accompanied by a binary label. '+1' denotes the positive class and '-1' indicates negative class. If every single vector holds m values, then the data is in m-dimensional universe which is known as the input space. The training algorithm of SVM checks the input space for a plane (or, when m > 3, a hyper - plane) that divides the positive and negative samples. Figure 10 shows the structure of SVM with hyperplane shown by the red line and support vectors by dots. According to learning theory, when there are several such hyperplanes, an optimal approach chooses the farthest hyperplane from every training set. This hyperplane is referred to as the largest margin hyperplane. The problem of picking the hyperplane of maximum margin for a given dataset can be defined and addressed quickly using quadratic programming. The training phase of SVM begins with this optimization [48].
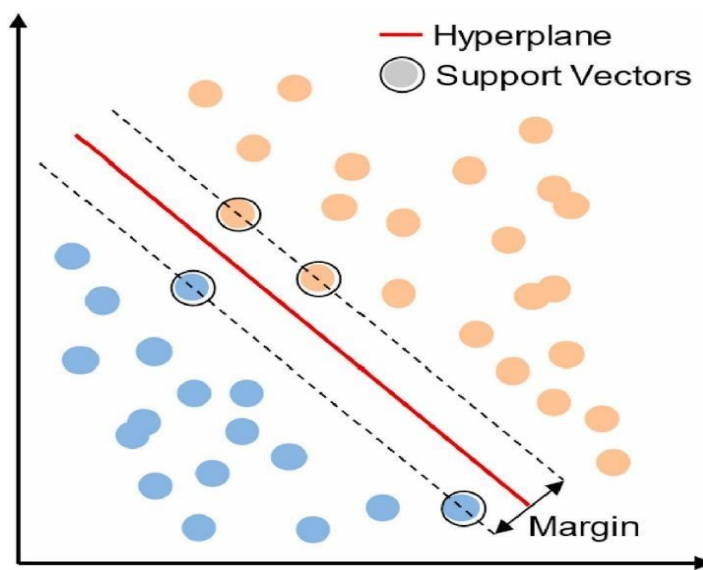
Figure 3.2: *Structure of SVM [48]*

This explanation of the SVM method is lacking in detail, but it should enough for our upcoming discussion. Particularly, it has not been demonstrated how does the SVM operates if there is no separating hyperplane. In a nutshell, this dilemma is addressed in two directions. The first approach includes incorporating a 'soft margin,' which permits a portion of training set to lie on the 'wrong' edge of the plane; for example, few samples

labelled '-1' may sit on top of the '+1' area and vice versa. The second method consists of adding a kernel function which is elaborated below.

3.7   Kernel Methods

The general kernel function explains why two items are similar. A large value indicates a similarity between x and y, whereas a small value indicates a distinctness. When using kernel approaches in machine learning, the kernel function needs to be symmetrical and positive semidefinite. The fundamental idea behind a kernel method is simple yet nuanced. Let us consider that we have a set of n vectors with m length. An m x n matrix can be used to convey this data. Using a kernel function K (), we may determine the proximity between every set of vectors in the database. The values of these kernels can alternatively be represented as a kernel matrix, which is a n x n matrix. to be a kernel method, an algorithm must be able to demonstrate that the kernels matrix is a suitable interpretation of the data. In other words, it should be able to remove the original matrix of the data while still running the algorithm with just the kernel matrix.

Kernel function is a dot product (Scaler product). $K\ x, y = \sum i\ x_i y_i$. A kernel procedure is thus a system that may be created down so that all data vectors fit inside a dot product process. Easily replace the dot product operation with the K in order to make the algorithm 'kernelize'. The kernel function can be substituted for the dot product operation since it is mathematically equal to the projection of the database into a different dimension. Let us suppose there are m dimensions in the input space, However, we employ a quadratic kernel function which is specified as $K(x, y) = (\sum i\ x_i y_i)2$. A kernelized form of the SVM determines the hyperplane of extreme space in the feature space by resolving the basic optimization problem with a new kernel function [49].

3.8   Ensemble Tree (ET)

By assembling a group of base learners and combining their output findings to make a final judgement, ensemble learning model is a subfield of machine learning that can help enhance the efficiency and precision of generic machine learning classifiers. Use the complimentary information from those various classifiers to improve the performance and

judgement accuracy of EL classifiers [50]. Ensemble approaches are divided into two categories. The first is boosting, which is frequently employed to enhance how well machine learning models perform. The first type of ensemble learners of this kind to be imposed is AdaBoost. The second is bagging, which is mainly used in analysis related to regression and classification.

3.9 Bagging

Bagging, often known as bootstrap aggregation, was one of the early ensemble techniques. During a portion of the training phase, each ensemble core classifier (decision tree or other classifier) is trained. To pick a training set for each base classifier, bootstrap sampling is utilised, which involves randomly selecting and replacing a part of the dataset. The simple majority voting method is used to categorise new instances. Each classifier classifies the instance, and the final classification is established by summing the classifier votes. Figure 11: Bagging ET Process Diagram. While bagging is one of the simplest and most rudimentary ensemble classifiers, it should be mentioned that it did really well in a few experiments.[45].
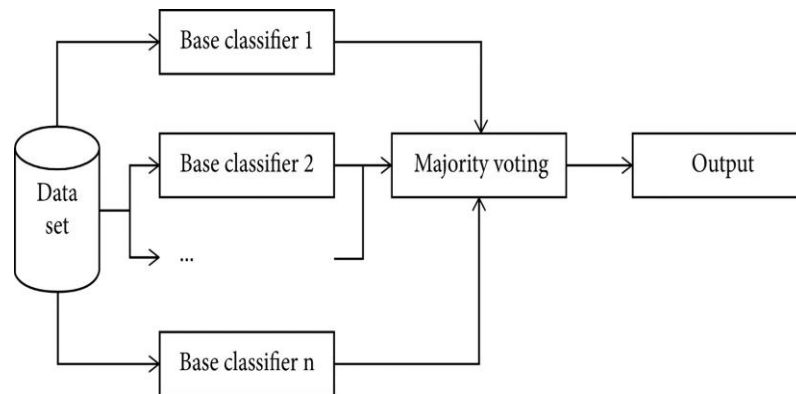
Figure 3.3: *Process of Bagging ET [45]*

Random forests are a bagging algorithm version in which the base classifiers are decision trees. Random forests, like bagging, employs bootstrap sampling and un - weighted committee aggregation for final classification. However, random forest method differs in the way decision trees are built. Typically, m features are chosen at random from

a set of M characteristics, and the ideal value of m is typically the square root of M. Furthermore, every tree in RF has not been pruned and is fully grown. Random forests perform exceptionally well, particularly for datasets with many features.

3.10 Boosting

Each base classifier is not entirely generated by the boosting approach by itself. Rather, the classifiers are created one after the other, with each base classifier that comes after giving the errors of the previous classifier more weights. Adaptive boosting is one of the most used boosting methods. AdaBoost uses weighted voting, data sampling, and an arbitrary number of base learners. First, the system builds a base classifier that is trained on the dataset using equal weights. Later classifiers are constructed by giving bigger weights to instances that the previous classifier misclassified, whereas examples

Figure 3.4: *Process of Boosting ET [51]*

successfully classified are given the same weights. After that, the weights including all instances in the entire dataset are normalized and applied for selection for the following classifier. Weighted base classifiers are used in the final classification. Figure 12 shows the boosting ET process [51].

3.11 Gaussian Process Regression Model (GPR)

GPR is a nonlinear, nonparametric regression model with a random attitude. This approach may measure forecast uncertainty in addi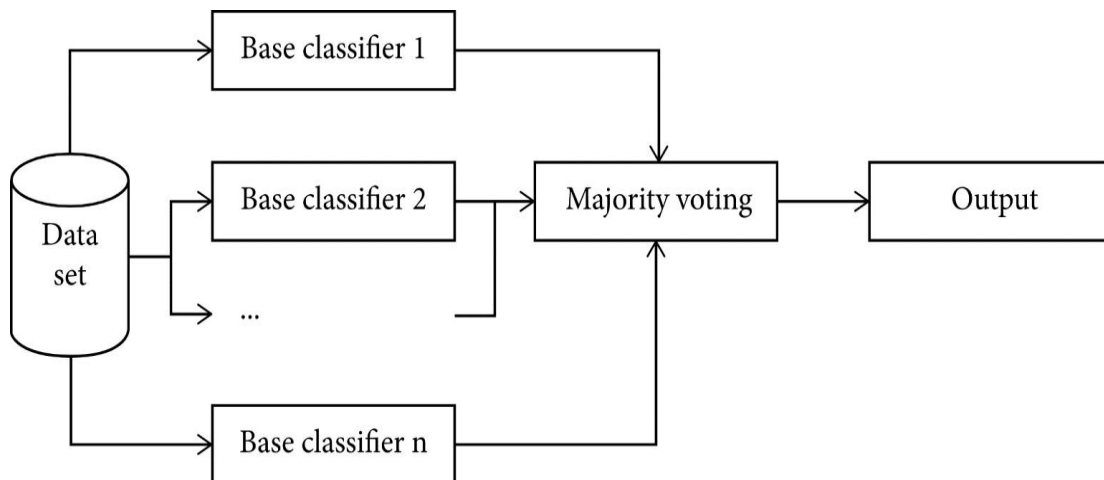tion to generating predictions by reporting the coefficient of determination for each prediction point [52]. To compute the probability of an input vector, utilise the Gaussian distribution's variance and mean. Rather of creating a variance and scaler mean, the GPR technique generates a mean and correlation vector. Using GPR, the smoothness proposal (represented by the interpolation zone) may be chosen quickly and quantitatively [53]. Figure 13 illustrates the basic workings of GPR.

.
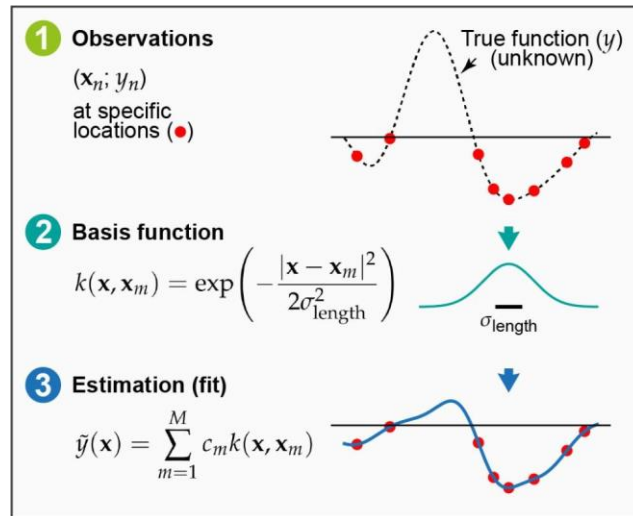
Figure 3.5: *Observations (3) Basic functions (3) Estimations [54]*

Let f represents a function (unknown) that converts inputs to outputs. $Y: f: x — y$. We indicate mathematically describing the relationship among input and output when we say we're modelling a function f. An accurate f model allows us to anticipate the output for a wide range of input values. This involves gathering observations of input and output and creating precise forecasts for freshly detected points based on these observations. For example, we will use GPR to simulate mouse trajectories in categorization research. Furthermore, we will use compositional GPR to deconstruct temporal dependencies in people's reaction times into intriguing patterns. To truly depict a function, one must actively choose the input points from which to examine the results. This process is known as

function exploration. The primary objective of basic exploration problems is to ascertain the underlying function as quickly and accurately as feasible by conducting a thorough investigation. Since this arrangement involves adaptively selecting the source locations based on prior knowledge about the variable and where information might be enhanced, it is closely related to optimal experimental design scenarios.

The goal is to locate inputs that yield maximum output in order to optimize reward acquired within a specific time frame. Exploration merely serves the objective of most effectively doing so. The purpose of this setup is to find the highest of the function as effectively as feasible, which is strongly related to design problems. It is known as exploration-exploitation because situations in which the yield of the underlying function must be optimized demand us to extract both ambiguous zones just to gain more information about exploration and input points that are likely to produce higher output given based on current knowledge of the exploitation.[54]

3.12   Overview of Optimization Algorithms

The selection of features and tuning of hyperparameters were done using two separate optimisation techniques, which were then compared. Particle Swarm Optimisation is the second, and Genetic Algorithm is the first.

3.13   Genetic Algorithm Technique (GAs)

General Analytic (GA) is a problem-solving technique inspired by natural selection. GA aims to assist a population in adjusting to shifting environmental circumstances. Because each person is different, good habits need to be refined while bad and unnecessary traits need to be lost. GA is an evolutionary algorithm that utilises programming and evolutionary techniques. The procedure was started with a population that was created at random [59]. The operators, which are explained below, are what make up the GA.

Many researchers have used GAs to evaluate the solution of complex issues involving decision variables that lack consistency, uniqueness, and other features. During searches, GA aggregates potential solutions into data structures resembling chromosomes

23

and keeps a community of these genes. It necessitates the employment of a fitness function that awards each solution a scalar reward (or incentive). If GA has created a good plan and evaluation function, they can expect honest responses. It then generates an initial solution of a predefined set of threads or chromosomes, known as that of the population size. The very next step is to analyze every option in the native population using a payout function. The rewards for superior responses are larger, while the rewards for the remaining options are lesser. The following generation is formed using GA operators including crossover and mutation. This practice is repeated till an ideal solution (or solutions) is found, the number of iterations or populations is reached, or the discrepancy among solutions is less than a predefined limit is attained. Figure 16 shows the schematics of GA. [60,61] The aspects of GA are described briefly below:
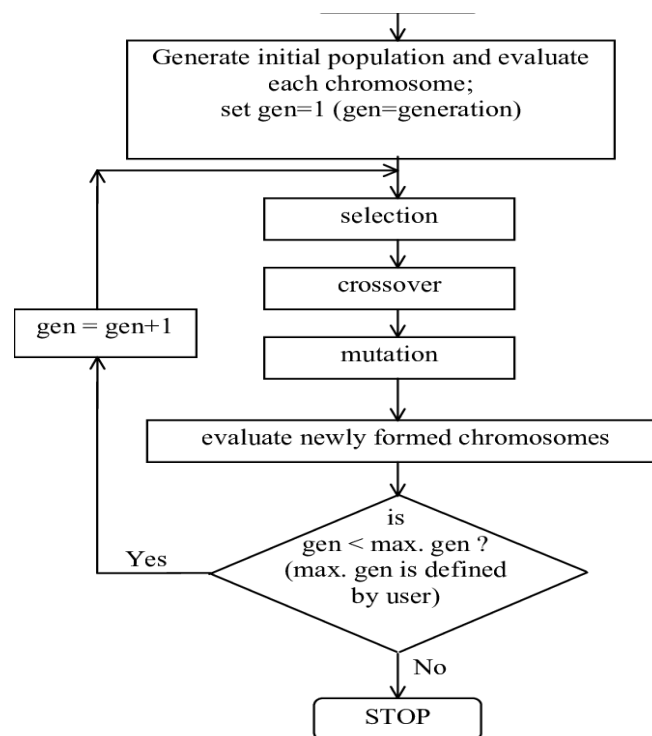
3.14   Representation

Figure 3.6: *Workflow of GA*

The genetic algorithm requires solutions or individuals in a population to be depicted by chromosomes. The representation scheme utilized determines the structure of

24

an issue and the sort of genetic operators which will be used. Specific alphabets are employed to create a gene sequence that makes up the chromosome. These alphabets can be made up of binary numbers (0 and 1) and value numbers. It has been demonstrated that chromosomes encoded with real value values yield more effective GAs and better solutions [62].

3.15   Selection Function

In GA, thousands of generations are created by selecting people from a preceding generation. Classification is founded on the idea that each and every entity has a chance or likelihood of being chosen once more than once for reproducing in the following generation, based on their fitness value. Some of the selection procedures include roulette wheel selection, scale techniques, competitions, elitist systems, and rating scales. A common stage in these approaches is assigning individuals a chance of selection. This task can be completed using a roulette wheel, logarithmic ranking, or geometrical priority.

3.16   Genetic Operators

GA operators alter the population over generations to achieve the desired or optimal result. The adaptive qualities from previous generations need to be maintained by the population. The operator in charge of selection searches for parents whose kids exhibit high levels of fitness. Chromosomes with better fitness values are chosen for reproduction. Chromosome selection and fitness value are closely associated in reproduction. Another GA operator called crossover causes parent traits to be integrated during reproduction, producing offspring with characteristics derived from both parents. Mutation operators are used to stop advanced individuals from entering populations and to stop important information from vanishing from the string. t randomly modifies one or even more individuals in a group to guarantee genetic diversity. Additionally, binary coded gene is generated by reversing the gene sequence from 0 to 1.

3.17   Initialization

GA requires an initial population to begin the process of determining the optimum solution. The population can be generated by producing random responses inside the

variables' upper and lower bounds. Another approach is to increase the starting population with well-established best approaches to improve on the existing ones. The rest of the population could be generated at random. Once a specific termination meets the criteria of termination then the GA operations are terminated. The requirement can be any or a mix of the following: (1) The number of generations approaches a predefined maximum value. (2) The population adapts on a unique solution. (3) The difference between solutions becomes less than a specified criterion. (4) The ideal solution does not grow over a certain no. of generations. (5) The evaluation values meet an acceptable level. [63]

When the goal function was not reached, GA was refreshed to create a new batch of population. First, a population was produced at random. A fitness function was then used to assess the features. GA operators were used to generate the population. Table 1 displays the GA parameters that were utilised to produce the optimised ML models in this work.

This method is repeated until the best fitting function is achieved. Table 1 reveals the parameters of GA used in this research.

Table 3.1: *Parameters of Genetic Algorithm*

GA Parameters	Generation	Crossover Probability	Crossover	Elite Count	Size Population	Type Population	Mutation	Probability of Mutation
Values	100	0.8	Scattered	3.95	50	Bitstring	Uniform	0.1

3.18 Particle Swarm Optimization Technique (PSO)

Fish schools and bird flocks are two examples of animals whose collective behaviours first had an impact on PSO. PSO solutions are displayed in the search box as "birds," also known as "particles." A cluster of these particles' combs over the search space in quest of the best location. In a multidimensional N space, the particle carries the location vector and the velocity vector. The multidimensional space's dimension is represented by $'j'$, while the vector's location is denoted by $'i'$. The velocity vector in this instance is defined as follows: $V_{ij} = (v_{i}1, x_{i}2, \dots, x_{i}N)$, where N denotes the number of unknown variables

26

or the space's dimension. In this instance, VL represents the VL location vector and is defined as VL = (VL1, VL2,, VL).

The two equations given below are used for updating the velocity and position of the particles, where k is iteration number used for updating both vectors.

$$V_{ijk} = wV_{ijk} + C_p \ r_p(X_{pbestijk} - X_{ijk}) + C_g \ r_g(X_{gbestijk} - X_{ijk})$$

$$X_{ijk+1} = X_{ijk} + V_{ijk+1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1)$$

When C_p, and, and C_g are positive reals, the optimal or best position of the swarm thus far is represented by $'X_{gbebij}'$, and the general ith personal best position of the particle is represented by $'X_{pbjj}'$. The inertia weight, denoted by $'w'$, represents the impact of the preceding velocity vector on the subsequent one, while $p \ and \ g$ are random values generated with a uniform distribution on a unit interval. In every dimension, the velocity has an upper bound V_{max}.

In this research the values of PSO parameters were [36]:

$$inertia \ weight = 1, \ social \ adjustment \ weight = 1.49, \ self -$$

$$adjustment \ weight = 1.49, min. \ adopted \ neighborhood = 0.25$$

CHAPTER 4: METHODOLOGY

4.1 Summary of Methodology

In this research, data for the liquefaction of animal and plant biomasses was taken and different ML models were developed integrated with optimization algorithms (1) PSO and (2) GA. The operational parameters include Ultimate, Proximate analysis, and liquefaction conditions.

Data was imported from an excel sheet in MATLAB R2021. ML models were combined with optimisation techniques, such as GA and PSO, to pick features and forecast the yield of bio-oil. In addition, the data set that forecasted the yield of ethanol from biomasses under fictitious conditions was used to train an AI system.

In this study, a GPR-PSO based model was developed to predict the bio-oil yield from the liquefaction of animal and plant biomass waste using Proximate and ultimate analysis and liquefaction parameters. The workflow of this research is sn in figure 17.
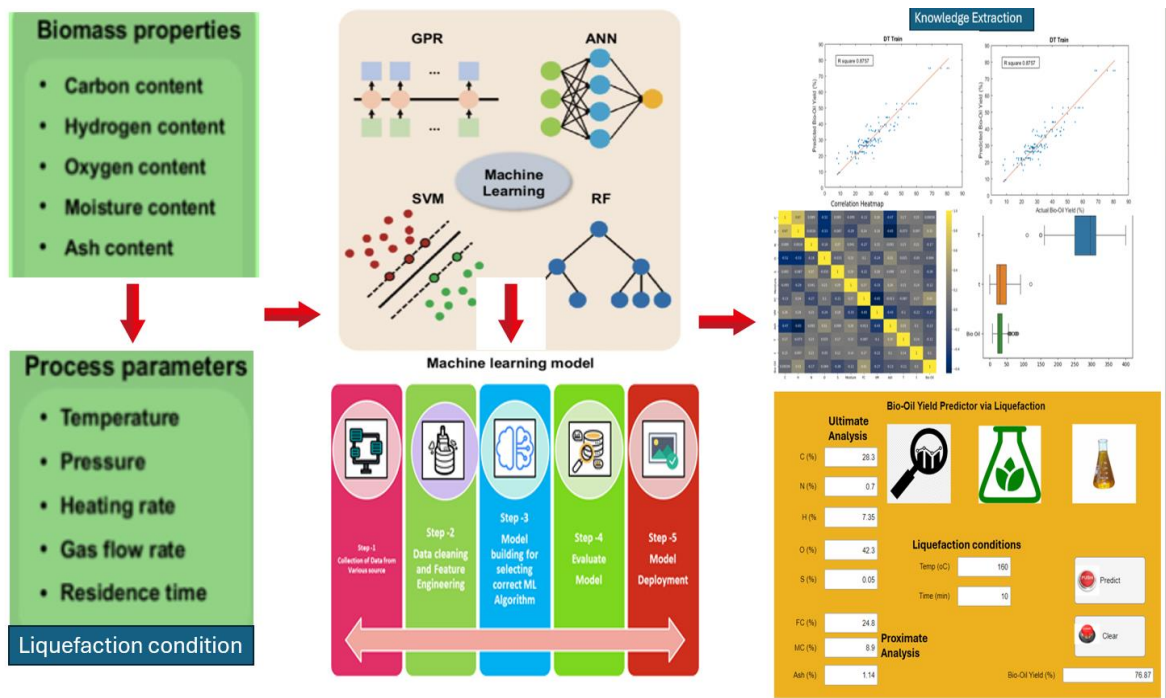

Figure 4.1: *Research Workflow*

4.2   Data collection

Data from the manufacture of bio-oil from the liquefaction of animal carcasses, barley straw, and pine wood are gathered for this study. To locate pertinent researching articles or articles in the databases of Google Scholars, Science Direct, and Web of Science, a variety of keywords were utilized, including pine wood, barley straw, animal carcass biomass, bio-oil, liquefaction, plant biomass, and machine learning. Division and feature data extraction were completed in order to assess and construct machine learning models following the collecting of data from pertinent articles. The procedure for gathering the data is explained in more depth below.

- A total of 312 bio-oil yield data sets were gathered from research papers.

- Samples for liquefaction data were gathered by considering various biomasses of animal carcass ultimate and proximate composition, pine wood, and barley straw, as well as liquefaction factors.

- Tables, supplemental data, and figures from the literature provided the information for the data sets.

- Information on various biomasses of pine wood, barley straw, and animal carcass ultimate and proximate composition—which includes moisture content, ash, fixed carbon, and volatile matter—was acquired from the literature. Carbon, oxygen, nitrogen, and hydrogen make up the ultimate composition.

- In the liquefaction of pine wood, barley straw, and animal carcass, input features such as liquefaction temperature (C), residence duration (min), and ultimate and proximal compositions were employed.

- MATLAB was used for the datasheet preprocessing, and various filling techniques (previous, nearest, next, linear, spline, pchip, and makima) were applied to fill in the missing data. The filing techniques that produced the fewest errors were employed. Various techniques for eliminating outliers were also applied.

- Bio-oil yield was the output of the liquefaction of pine wood, barley straw, Animal Carcass biomasses.

- Tables 1, 2, and 3 display the entire data distribution.

- Using the ultimate, proximate analysis of animal and plant biomasses and the conditions of liquefaction, a software with a graphical user interface was developed to forecast the generation of bio-oil. To develop the GUI, the MATLAB R2023 toolbox was utilised in future.

CHAPTER 5:   RESULTS AND DISCUSSION

Following up on the previous line of the literature review, this study looked at five different machine learning techniques for predicting the yield of bio-oil based models: SVM, GPR, Ensembled Tree, DT, and ANN. Figure 18 illustrates how ML models use the composition of plant and animal biomass as well as the conditions for liquefaction as input parameters of liquefaction process. The biomass composition consists of Ultimate analysis which includes the percent concentration of Carbon (C), Hydrogen (H), Nitrogen (N), and Oxygen (O) and proximate composition, which is the percent concentration of Ash, Fixed Carbon (FC), Volatile Contents, and Moisture Content. The Liquefaction parameters include, Liquefaction Temperature (T), and residence time (t). The product of liquefaction is bio-oil.

This study used liquefaction characteristics (temperature, time, and bio-oil yield) along with proximal and ultimate analyses (biomasses composition) to develop machine learning (ML) models, including SVM, ANN, DT, GPR, and ET, to forecast the bio-oil output from biomasses liquefaction. This research is novel in that, instead of using time-consuming and potentially overfitting filter and wrapper approaches, two different optimisation methods—PSO and GA-based algorithms—were used for features selection and ML model hyperparameter optimisation and compared. While the PSO-based features selection applied in this study is an advanced and novel method of optimization that uses natural ways to solve problems. It takes less time and optimizes results depending on cross-validation performance.

PSO and GA are both evolutionary search techniques, which means that they shift from one set of points to another inside an iteration with evident improvement over the prior values via deterministic and probabilistic principles. After features selection five different machine learning models were trained and established to compare their prediction performance of bio-oil yield (Figure 18). Additionally, a study of partial dependence, which shows how input features affect target variables, and the significance of input

variable features were carried out. Apart from that correlation plot for the relation between any two variables was constructed.

Finally, an extremely simple software tool for calculating bio-oil yield was built based on the suggested GPR model. The complete understanding of microalgae pyrolysis as well as practical recommendations were given by the anticipated accuracy of machine learning models and fresh perspectives that accompanied them.
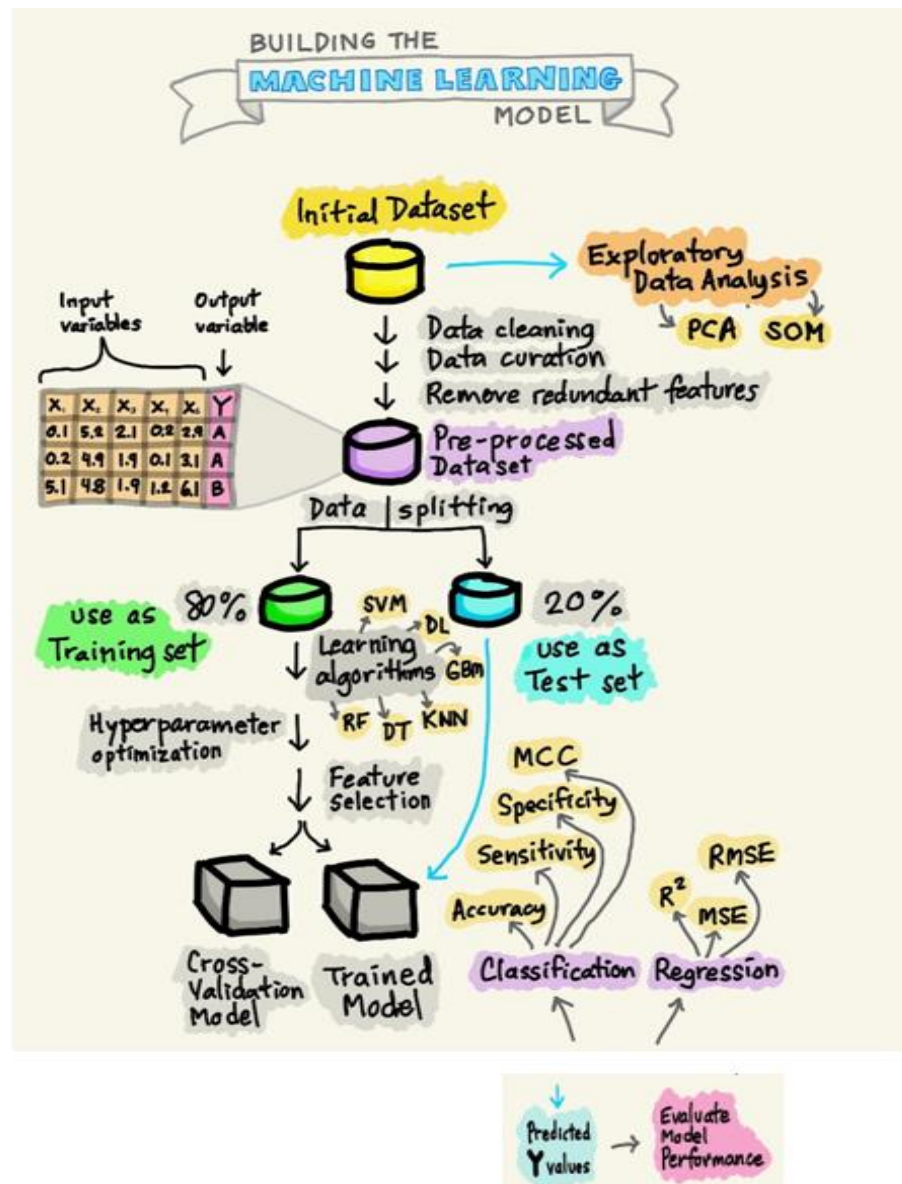

Figure 5.1: *Building of Machine Learning Model*

5.1 Box Plot Representation of Oak Tree and Animal Carcass Composition and liquefaction Conditions

The box plot, also known as the box-and-whisker plot, depicts the descriptive statistic and the first data division. Thus, the box plot allows you to visualise the peak, upper quartile, average, lower quartile, and bottom of any data set. As shown in Figures 19, 20, and 21, the parameters of the ultimate and proximate compositions, as well as the liquefaction conditions, were plotted against their distribution in the box plot.

The proximal analysis, which included FC, Ash, MC, and VM, yielded 4.1%-24.80%, 0.20%-59.93%, 0.17%-2.24%, and 47.20%-95.06%, respectively. However, the final study considers the following: 2.75-77.2%, 8.6-65.20, 3.01-13.88%, 0.1%-12.80%, and 0.05%-2.24% for carbon, oxygen, hydrogen, nitrogen, and sulphur, respectively. In addition, the liquefaction reaction parameters T, time, and bio-oil yield are 110%-400%, 0%-12%, and 7.8%-81%, respectively. Figure 20 displays box plots as depicted.
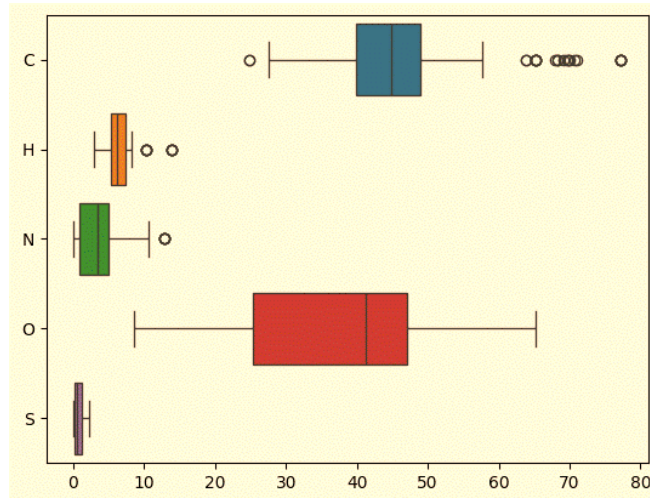

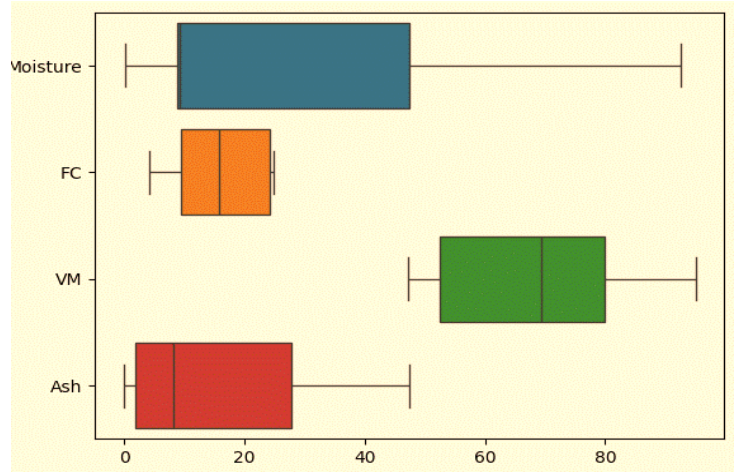**Figure 5.2:** *Box Plot of Ultimate Analysis*
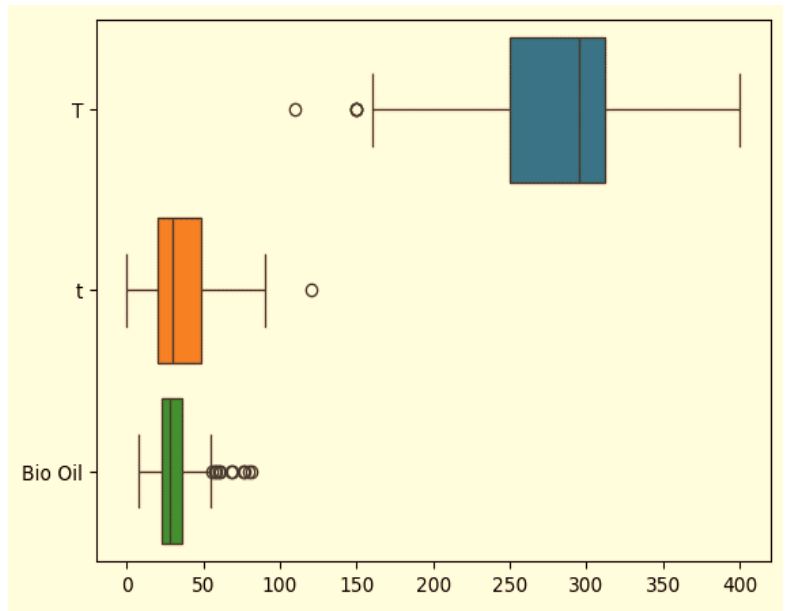
Figure 5.3: *Box plot of Proximate Analysis of Biomasses*

Figure 5.4: *Box Plot of Bio-oil yield Conditions*

Table 5.1: *Data Distribution of Ultimate Analysis*

Parameters	Min	Q1	Median	Mean	Q3	Max
Carbon	24.75	39.8	44.93	44.5	48.9	77.2
Nitrogen	0.1	0.8	3.44	3.8	5.0	12.80
Hydrogen	3.01	5.30	6.22	6.36	7.30	13.88
Oxygen	8.6	25.30	41.3	39.04	47.13	65.20
Sulphur	0.05	0.18	0.47	0.64	1.24	2.24

Table 5.2: *Data Distribution of Proximate Analysis*

Parameters	Min	Q1	Median	Mean	Q3	Max
FC	4.11	9.47	15.78	15.82	24.20	24.80
VM	47.20	52.50	69.40	69.41	79.8	95.06
MC	0.17	8.8	9.31	27.3	47.5	2.24
Ash	0.20	7.20	9.50	12.49	15.36	59.93

Table 5.3: *Data Distribution of Liquefaction Conditions*

Parameters	Min	Q1	Median	Mean	Q3	Max
Temp	110	250	295	281.05	312.3	400
Time	0	20	30	33.8	48.75	120
Bio-oil yield	7.8	22.8	28.17	30.9	35.7	81.00

5.2   Features Selection

Using an optimisation approach, features selection is the process of selecting the process's best features with the largest impact [1]. This study used two unique optimisation strategies to choose the characteristics. PSO is first, while GA is second. All machine learning techniques relied on these two optimisation methodologies. We picked features that had a major impact on biofuel yield. Using feature selection improved the R2 and RMSE values. The main principle behind feature selection is to improve the prediction ability of machine learning models.

Table 2 shows the features selected for ANN, GPR, SVM, DT, and ET through genetic algorithm and particle swarm optimization of the optimization techniques.

Table 5.4: *Features Selected for GA and PSO*

Model Type	Features Selected for GA	Features Selected for PSO
Support Vector Machine	C, H, N, O, S, MC, FC, VM, T, t	C, H, S, Ash, MC,FC, Ash
Gradient Process Regression	N, O, MC, FC, Ash,t	H, N, O, MC, FC, Ash, T, t
Ensembled Tree	C, H, O, S, Ash, MC	C, H, N, O, S, MC, FC, Ash, T, t
Decision Tree	S, O, N, Ash, FC, VC,	C, N, O, S, FC, VM, T, t

5.3   Performance Evaluation Criteria

GPR, ET, ANN, DT, and SVM were pre-processed from the MATLAB toolbox using default hyperparameters. Based on R2 and RMSE, the pre-processing techniques were assessed. Two categories were created out of the datasets: training (80%) and testing (20%). To reduce data waste and overfitting, 5-fold cross validation was used. After noting

the hyperparameter values *from the regression model toolbox, PSO and GA were used to optimise the models, and the results were compared. These hyperparameters were then employed in the development and testing of models. The evaluation of machine learning models was conducted using two key performance metrics, namely root-mean-square error (RMSE) and coefficient of determination (R2). The R2 and RMSE formulas are given below.*

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i^{exp} - Y_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i^{exp} - Y_i)^2}{\sum_{i}^{n} (Y_i^{exp} - Y_i^{exp})^2}$$

where Yi is predicted data, $Yexp$ is experimental value, n denotes number of tests samples.

5.4   Optimization of ML Methods

The regression model's toolbox contained the tuning parameters for each machine learning model, which were chosen using 5-fold cross-validation and standardised data. For GPR, SVM, ET, and decision trees, they were trained and optimised with the aid of PSO and GA. The kernel function "gaussian," kernel scale 2.8339, box constraint 765.5795, and epsilon 11.2812 were the optimised settings of the SVM hyperparameters in the instance of GA. In contrast, the values of the hyperparameters for PSO were Epsilon 6.2225, Box Constraint 198.0083, Kernel Scale 1.6970, and Kernel Function "gaussian." Comparably, in the ET scenario, the parameters derived by the GA optimisation were Method "Bag," Number of learning cycles (10), and Learning Rate (1). In contrast, the number of learning cycles for the PSO method "LSBoost" is 0. 7580.. For DT surrogate was selected 'on' for GA and 'on' for PSO and the minimum leaf size was equal to 1 for both. For GPR, in case of GA the value of Sigma was 7.2841, and for PSO Sigma was 0.9957.

5.5   Prediction Performance

Oak tree bio-oi output was predicted using machine learning models, namely SVM, GPR, DT, and ET. Each ML model predicted the bio-oil yield accurately based on selection of GA and PSO variables. Tables 3 and 4, respectively, present a comparison of all four models for feature selection based on PSO and GA. The GPR's R2 and RMSE values prior to feature selection were, respectively, 0.9309 and 0.1119. The GPR model's R2 value climbed to 0.5187 and the RMSE value dropped to 0.3852 after employing PSO for feature selection and optimised hyperparameters. In contrast, the GA-based features selection method yielded a higher R2 value (0.9270) and a lower RMSE value of 0. 1077.The performance of ET, DT, and GPR was acceptable for both GA and PSO as compared to SVM for bio-oil yield prediction as can be seen in figure 22 and figure 23 respectively.

Similarly For GA based study, the training and testing values of R2 for ANN, GPR, ET, SVM, and Dt were 0.9326, 0.9943, 0.9493, 0.5568, 0.62 and 0.9037, 0.8655, 0.9715, 0.9192 respectively as can be seen in figure. The root mean square error (RMSE) and coefficient of performance (R2) values were used to compare the performance of each machine learning model. When it came to PSO and GA-based feature selection, the SVM model performed poorly, whereas the GPR model performed well in all scenarios. When it came to PSO, the testing and training performance trend for every ML model was GPR>ANN>DT> ET >SVM. In contrast, every ML model in the GA-based study performed in the following order: GPR>ET>ANN>DT>SVM. Table 5 presents a comparison between the yield predicted by GPR and experimental data. The performance all the ML models were compared on the bases of the values of root mean square error (RMSE) and the coefficient of performance (R2). The performance of SVM was poor in case of both PSO and GA based feature selection while that of GPR model was best for both PSO and GA. In case of PSO, the performance trend for testing and training for all the ML model was GPR>ANN>DT> ET >SVM. While in case of GA based study the performance of all the ML model was in the order of GPR>ET>ANN>DT>SVM. The comparison of the yield predicted by GPR and experimental can be seen in Table 5.
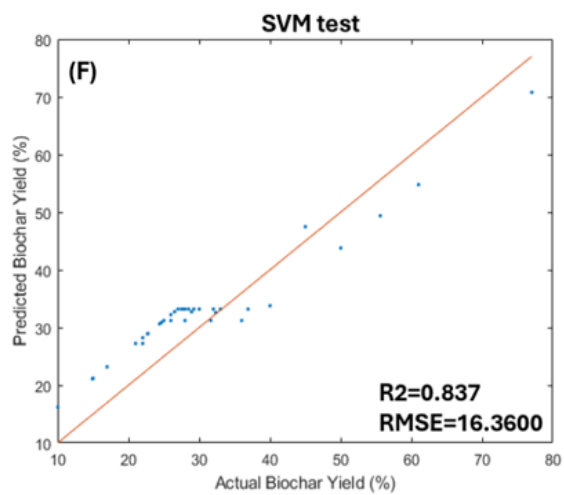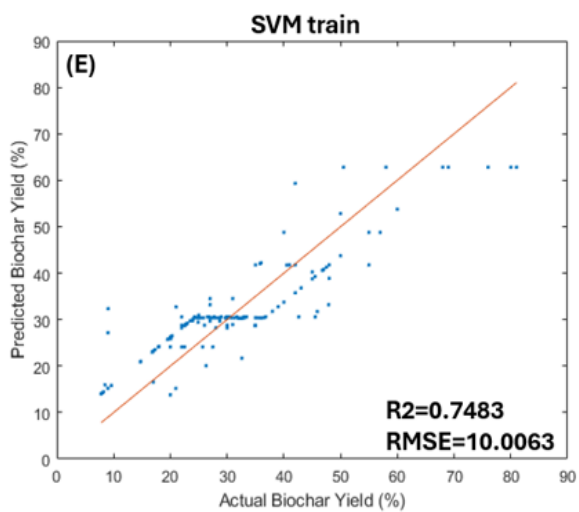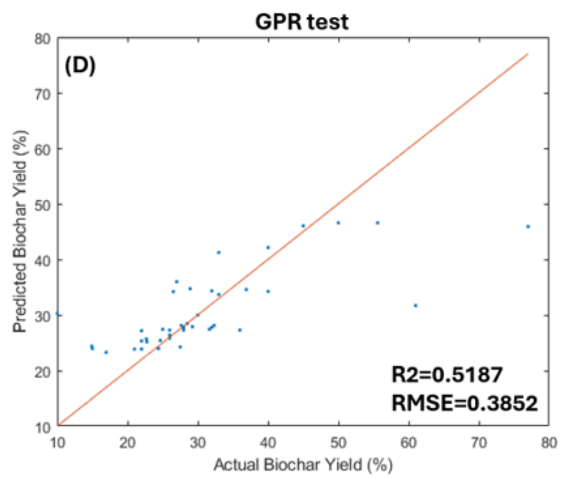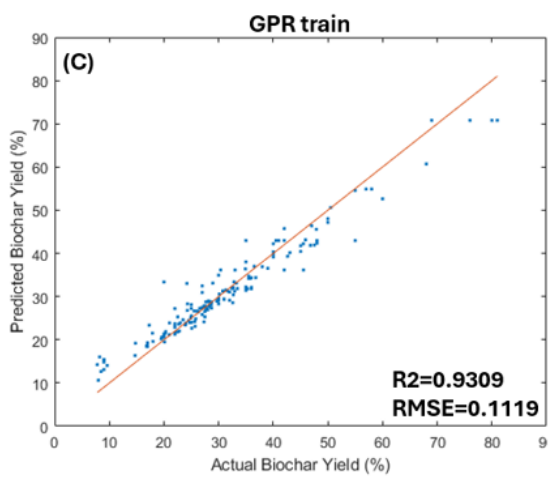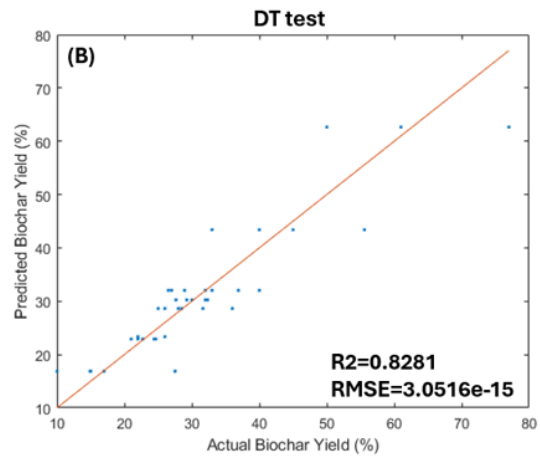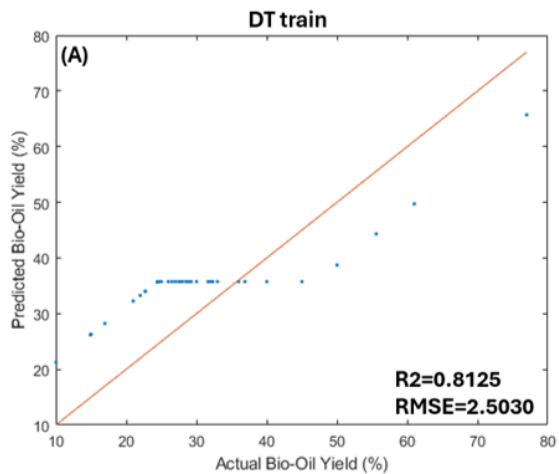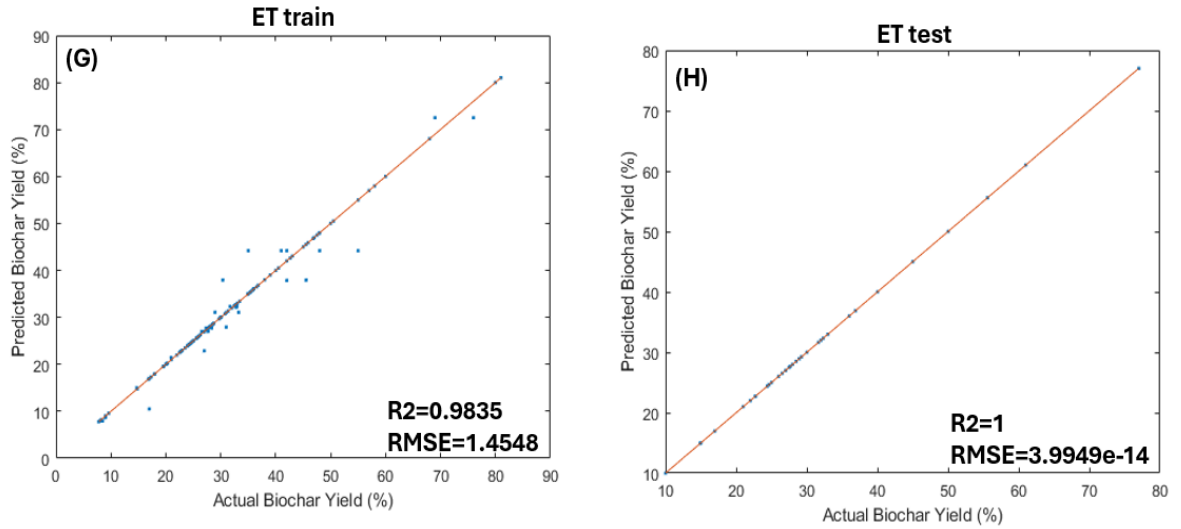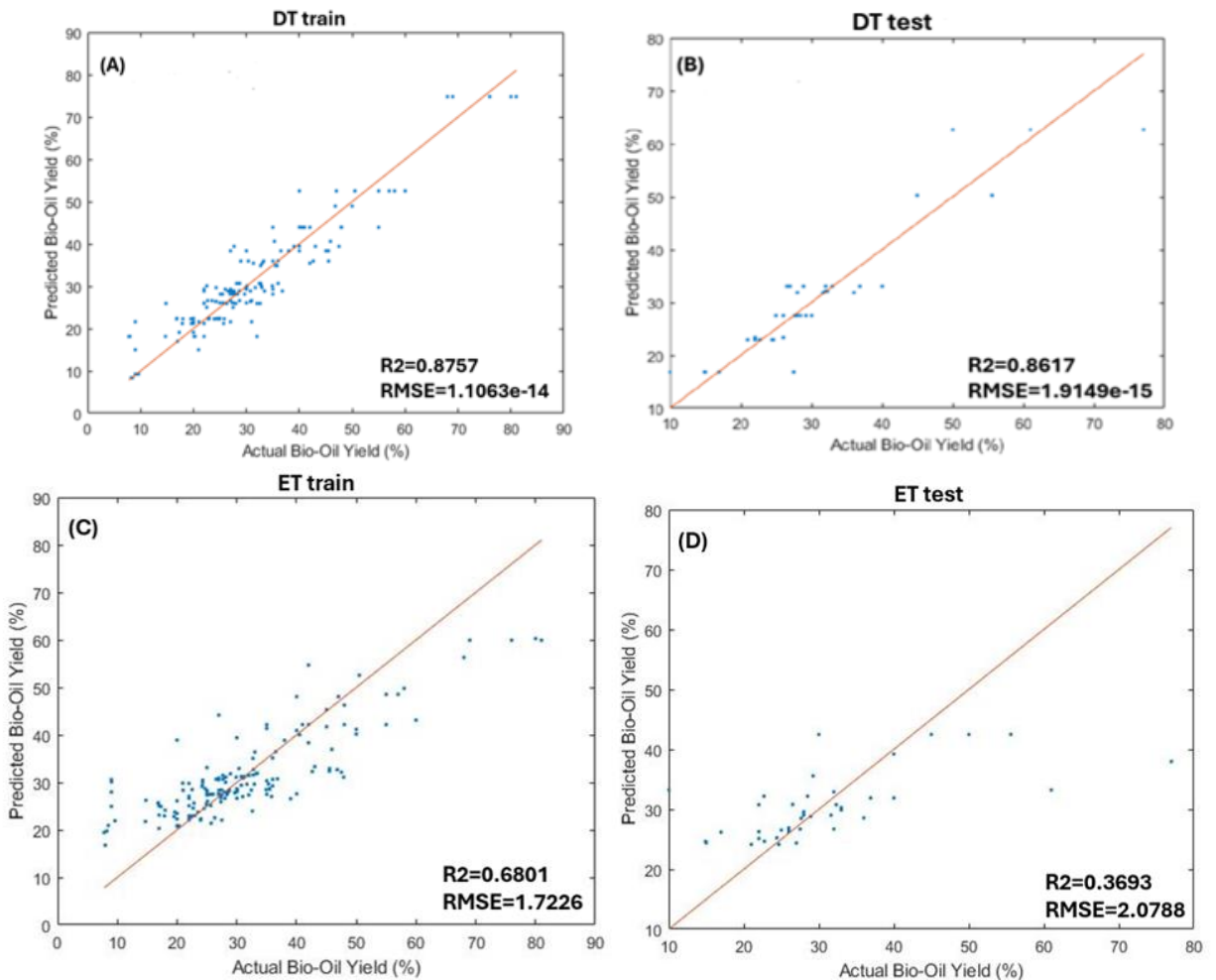
Figure 5.5: *Comparison of various ML models with bio-oil production (AB) DT, (CD) GPR, (EF) SVM, and (GH) ET, using PSO-based algorithm.*
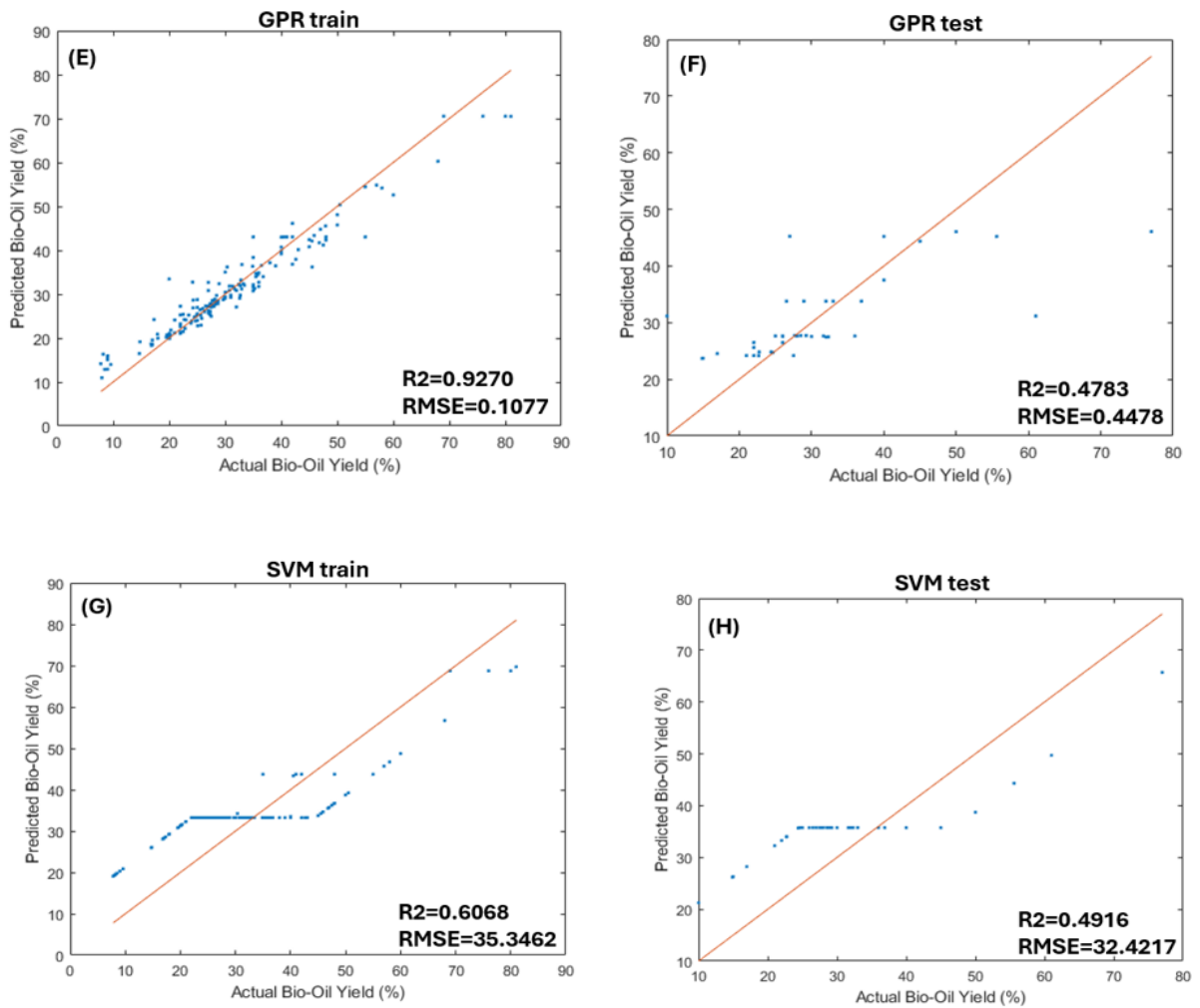
Figure 5.6: Comparing different machine learning models with the generation of biofuel (AB) DT, (CD) GPR, (EF) SVM, and (GH) ET, using GA-based algorithms.

Table 5.5: Comparison of ML methods using PSO.

Model	Training R^2	Training RMSE	Testing R^2	Testing RMSE
GPR	0.9309	0.1119	0.5187	0.3852
ET	0.9835	1.4548	1	3.9949e-14
DT	0.8125	2.5030e-14	0.8281	3.0516e-15
SVM	0.7483	10.0063	0.8237	16.3600

Table 5.6: Comparison of ML methods using GA

Model	Training R^2	Training RMSE	Testing R^2	Testing RMSE
GPR	0.9270	0.1077	0.4783	0.4478
ET	0.6801	1.7226	0.3693	2.0788
DT	0.8757	1.1063e-14	0.8617	1.9419e-15
SVM	0.6068	35.3462	0.4916	32.4217

5.6 Features Importance

The Shapley technique was used to determine the relevance of each characteristic. The GPR-PSO model accurately predicts the relationship between microalgal composition, pyrolysis conditions, and bio-oil production. The GPR-PSO model was used to assess the relative impacts of temperature, heating rate, time, volatile matter, carbon, nitrogen, oxygen, and moisture content. Figure 24 shows that while N2, O2, and C concentrations had a little effect on bio-oil production, temperature, time, and moisture content had a considerable influence. Furthermore, neither volatile matter nor heating rate had a substantial effect on yield.
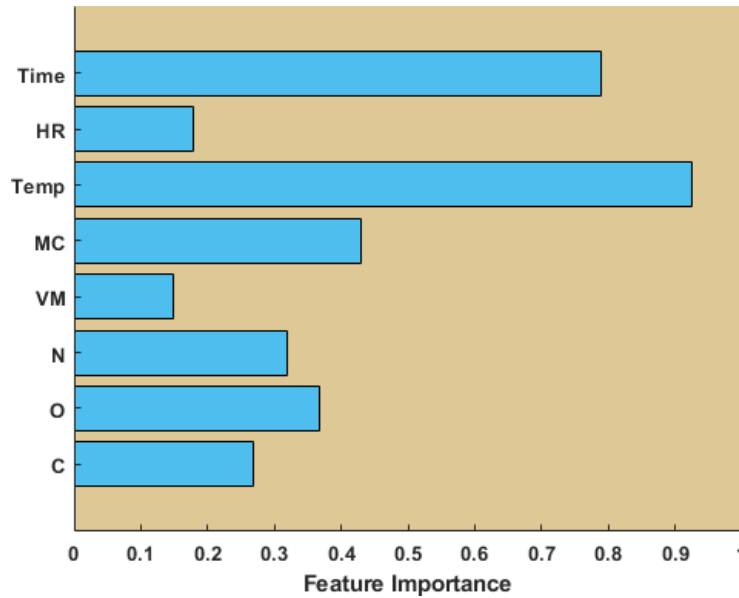
Trends followed is study is T>time>MC>O>N>C>HR>VM.

Figure 5.7: *Feature importance to produce bio-oil yield from the liquefaction of Oak Tree.*

5.7   Partial Dependence Plots (PDP) for Liquefaction Parameters

Plots of partial dependence were created for the feature chosen for particle swarm optimisation based on the liquefaction circumstances. Figure 25(A, B, C, D, E, F, G, H, I, J) displays PDPs for the proximate, ultimate analysis, and liquefaction conditions for the prediction of bio-oil produced. The relationship between carbon and hydrogen with bio-oil is depicted in Figures 25(A) and (B), respectively. According to [50], it is evident that adding more carbon and oxygen will boost the production of biofuel. Figure 25(C) illustrates how the nitrogen composition affects the yield of bio-oil. The most crucial factor in the liquefaction of oak tree yield is temperature, which also has a big impact on the quality of the bio-oil. They found that raising the temperature of the biofuel will result in a higher production of biofuel. As the temperature rises, more light ends are generated through thermal oil degradation, which causes the output of bio-oil to steadily fall. They explained the 400 °C increase in bio-oil yield by the synthesis of the main oil components at that temperature range. At higher temperatures, bio-oil with a higher carbon and hydrogen content was created, but the amount of bio-oil produced was less. Reactor

residence time is the second crucial factor in determining the yield of bio-oil. Less time spent in residence will be needed when the heating rate is increased.
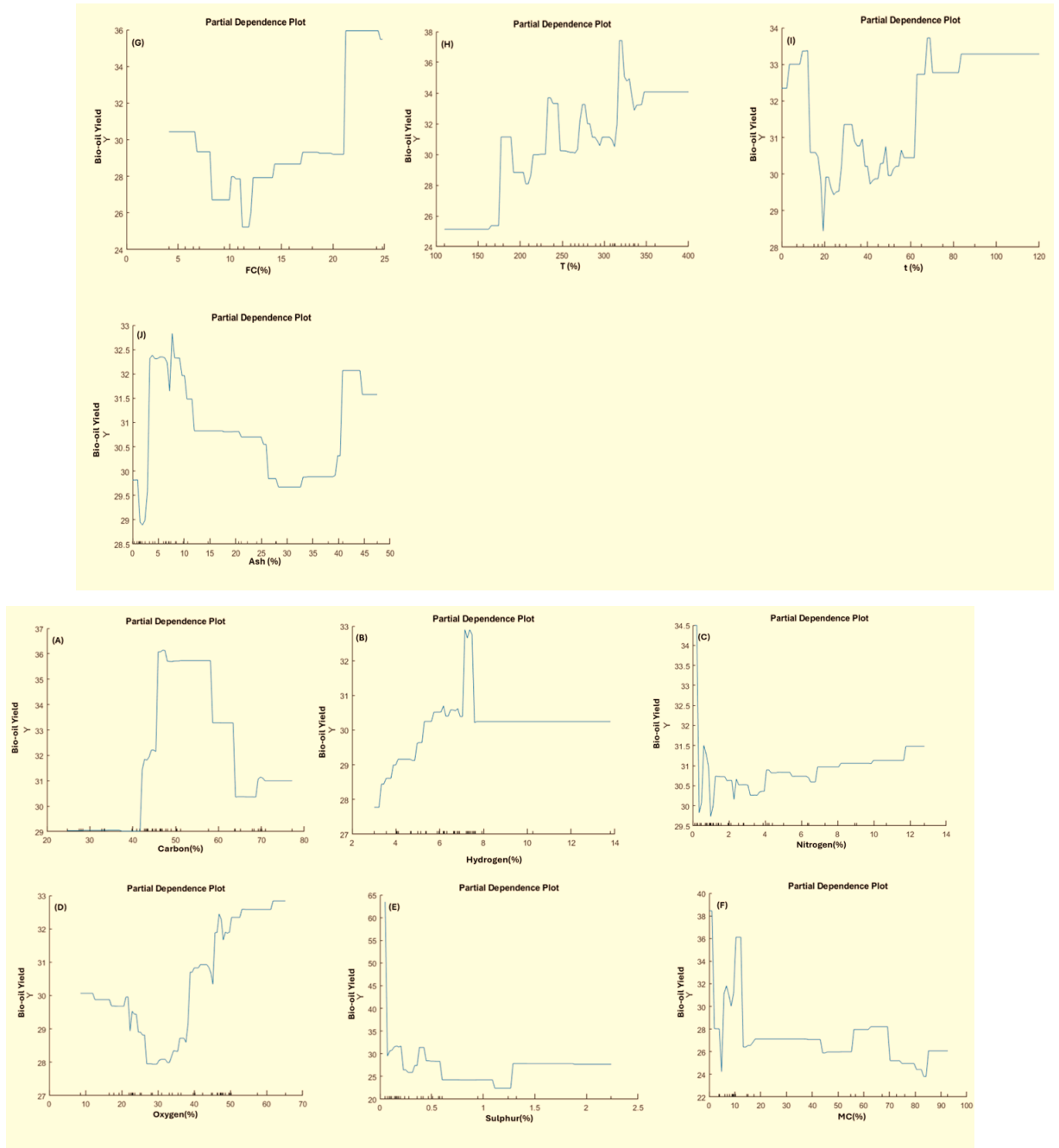
Figure 5.8: *Partial dependence plot for bio-oil Yield using Proximate and Ultimate analysis and Liquefaction conditions.*

A sample produces more liquid products and volatiles when heated quickly to an intermediate temperature, while prolonged heating produces more char at the same temperature.

Figure 26 displays three-dimensional plots that depict temperature and bio-oil production along with other significant features (time, FC, S, and C).

Figure 5.9: *The Partial Dependence Plot of Bio-oil produced on Temperature with (A) Time, (B) Fixed Carbon, (C) Carbon, and (D) Sulphur*

5.8   Correlation Plot

Figure 27 shows a graph of the Pearson correlation coefficient for any two parameters. The majority of the characteristics have non-linear interactions with one another, therefore they may be utilised to create models. O2, on the other han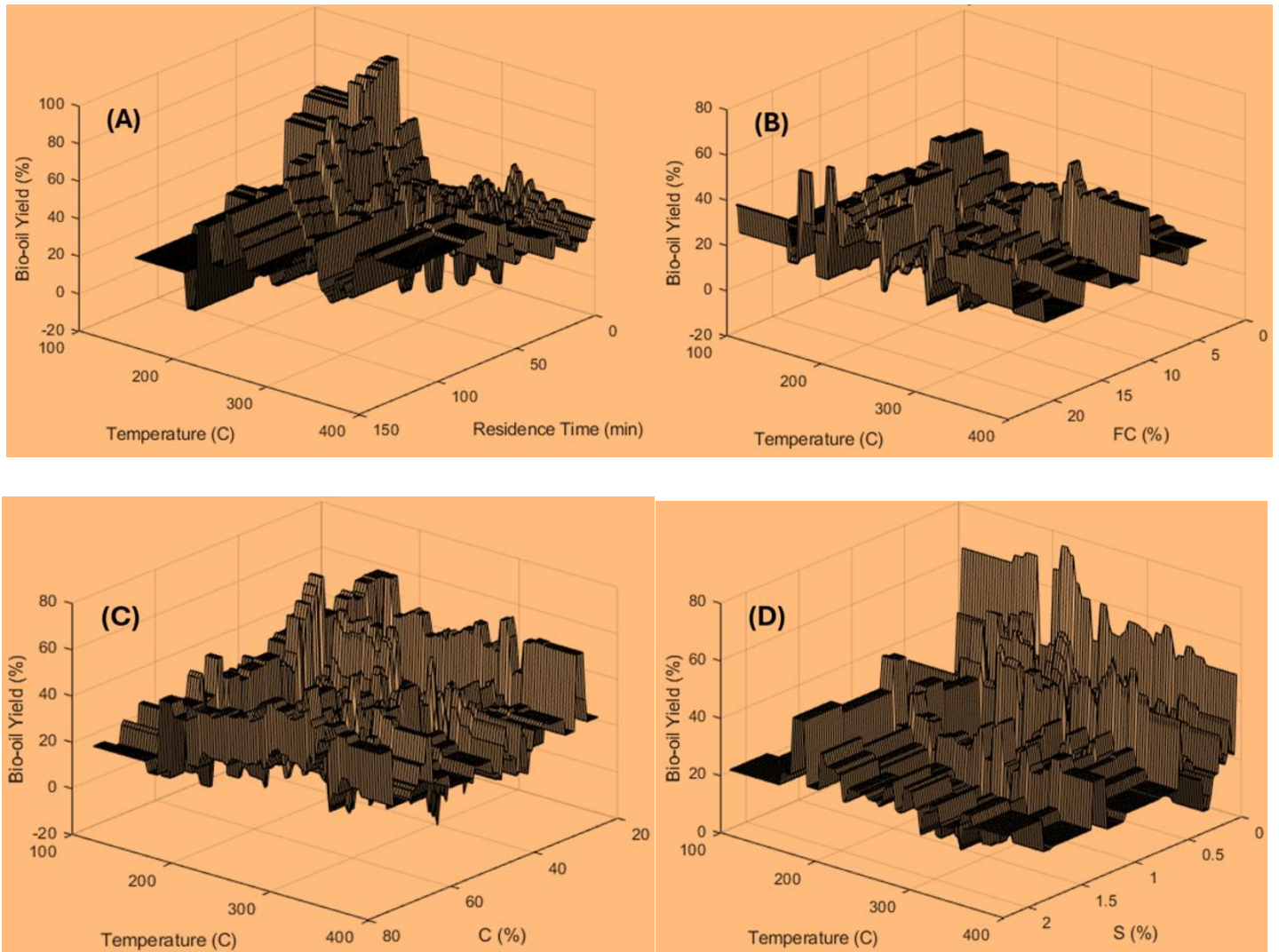d, has a very unfavourable association since the oxygen content is calculated using O = 100-C-H-N in the bulk of study. Similarly, FC, which is derived from the difference (FC = 100-Ash-VM), has a negative connection with VM, FC, and Ash. Furthermore, a significant positive linear correlation between C and H is seen, which is compatible with the number of C-H bonds in biomass feedstock. As a result, the inputs for hydrogen content (H) and FC must be eliminated from the following construction models.
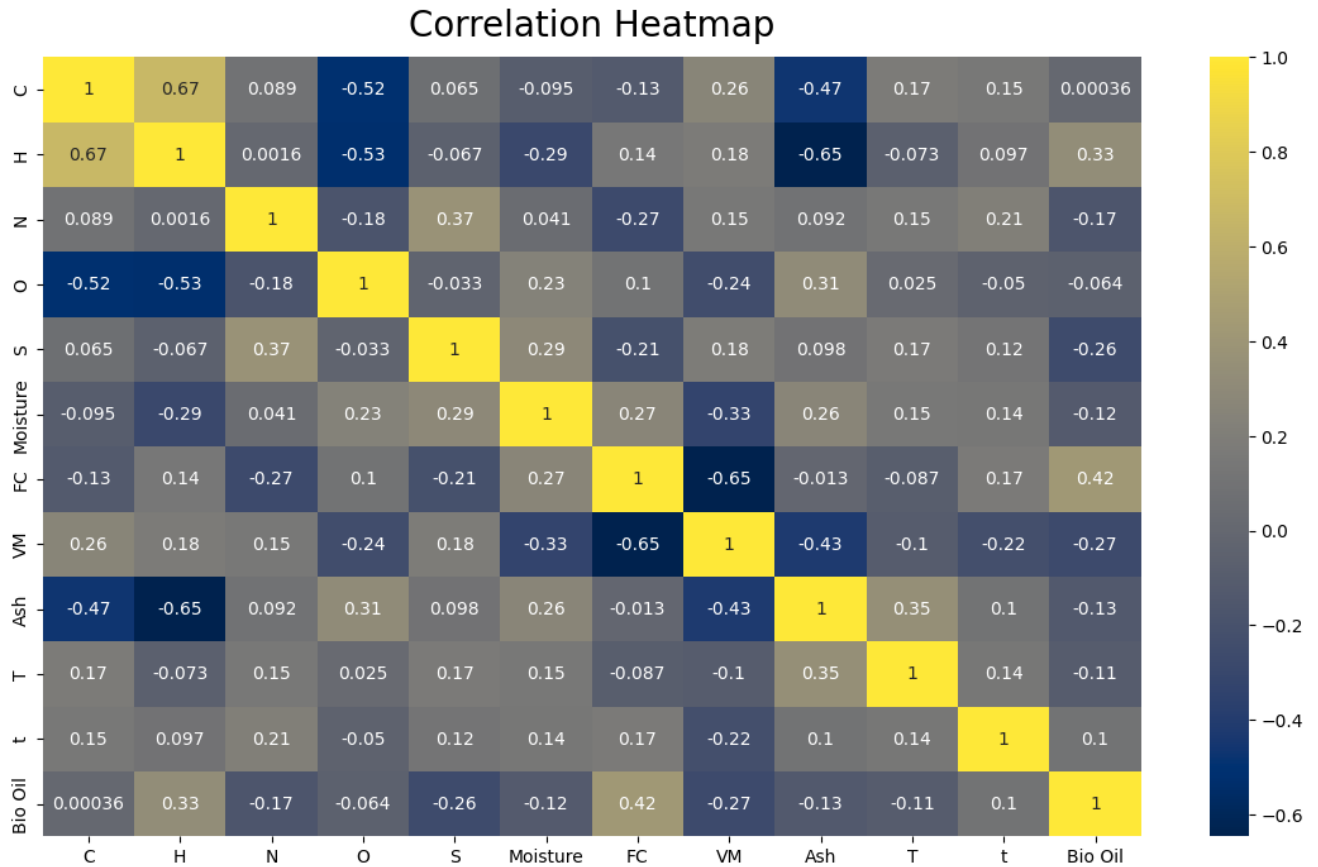
Figure 5.10: *Pearson Correlation Coefficient graph for any two variables*

5.9   Graphical User Interface (GUI)

Users can interact with electronic devices through the use of graphical icons, user-friendly applications, and symbols in a Graphical User Interface. Researchers were able to enter data about the liquefaction conditions, final composition, and proximate composition of microalgae using the GUI that was developed for this study. ET-PSO is the model prediction function that GUI employed in this work to forecast the yield of bio-oil. The GUI was developed using MATLAB 2021b. The GUI for oak tree composition and liquefaction parameters is shown in Figure 28. It includes the following: liquefaction conditions (Temp(C) = 160, Time(min) = 10); proximate analysis (FC=24.8%, MC=8.9% and Ash=1.14%); and ultimate analysis (C=28.3%, H=7.35%, N=0.7%, O=42.3%, S=0.05%).

Figure 5.11: *GUI for Oak tree feed and its respective bio-oil yield*

CHAPTER 6: CONCLUSION AND FUTURE

RECOMMENDATIONS

6.1   Conclusions

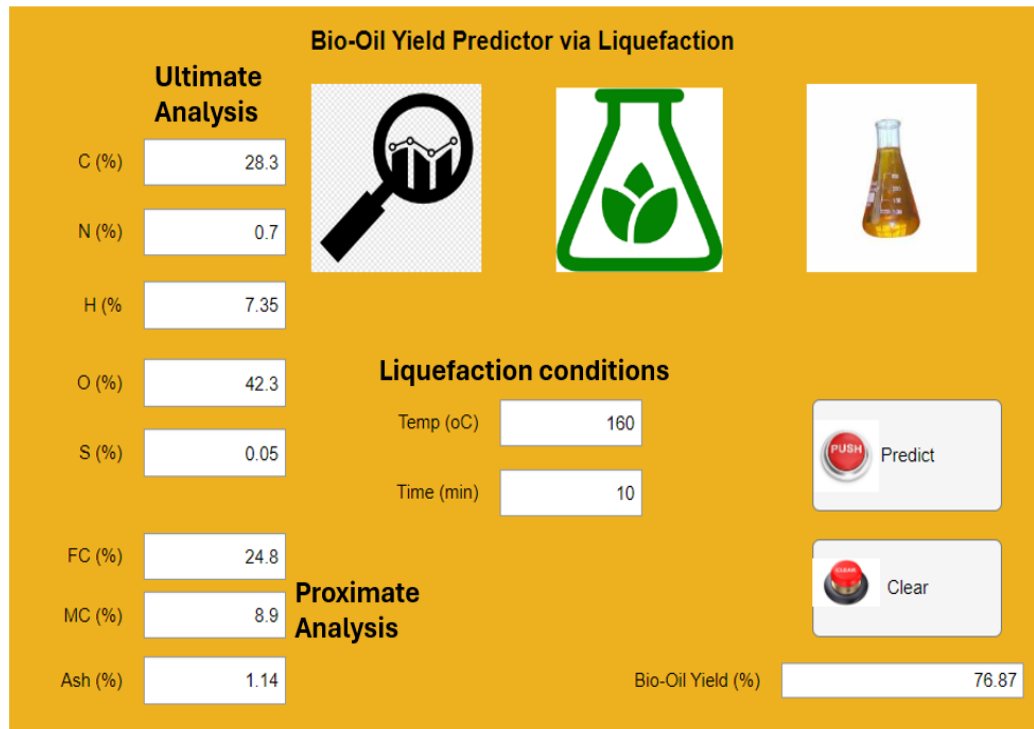This study employed four distinct types of machine learning models to anticipate the bio-oil output from the liquefaction of an oak tree. Using PSO for feature selection and ML model parameter optimisation, the ET model beat the other models in predicting bio-oil (R^2 = 0.9835) when compared to DT (R^2 = 0.8125), GPR (R^2 = 0.9309), SVM (R^2 = 0.7483), and SVM (R^2 = 0.43). The ET performed better while utilising GA for feature selection and hyperparameter optimisation. Furthermore, the partial dependency plot analysis demonstrated that the optimal parameters selected using the PSO-based technique had a substantial influence on the percentage of bio-oil in the output.

6.2   Future Recommendations

- A novel hybrid of GA and PSO can be applied in biomass liquefaction in future whichwill eliminate the cons of both the algorithms.
- The application of Deep learning can be useful in this field of Oak tree biomass liquefaction because it has never been done before.

REFERENCES

[1]   A. Boretti, "Hydrogen key technology to cover the energy storage needs of NEOM City," Int. J. Hydrog. Energy, vol. 47, no. 31, pp. 14371–14374, Apr. 2022, doi: 10.1016/j.ijhydene.2022.02.240.

[2]   A. R. Nia, A. Awasthi, and N. Bhuiyan, "Industry 4.0 and demand forecasting of the energy supply chain: A literature review," Comput. Ind. Eng., vol. 154, p. 107128, 2021.

[3]   C. Rao, Y. Zhang, J. Wen, X. Xiao, and M. Goh, "Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model," Energy, vol. 263, p. 125955, Jan. 2023, doi: 10.1016/j.energy.2022.125955.

[4]   A. Yahaya, N. M. Nor, M. S. Habibullah, J. Abd. Ghani, and Z. M. Noor, "Howrelevant is environmental quality to per capita health expenditures? Empirical evidence from panel of developing countries," Springer Plus, vol. 5, no. 1, p. 925, Dec. 2016, doi: 10.1186/s40064-016-2505-x.

[5]   J. M. Chen, "Carbon neutrality: Toward a sustainable future," The Innovation, vol. 2, no. 3, p. 100127, Aug. 2021, doi: 10.1016/j.xinn.2021.100127.

[6]   C. Zou et al., "The role of new energy in carbon neutral," Pet. Explor. Dev., vol. 48, no. 2, pp. 480–491, Apr. 2021, doi: 10.1016/S1876-3804(21)60039-3.

[7]   L. De La Peña, R. Guo, X. Cao, X. Ni, and W. Zhang, "Accelerating the energy transition to achieve carbon neutrality," Resour. Conserv. Recycl., vol. 177, p. 105957, Feb. 2022, doi: 10.1016/j.resconrec.2021.105957.

[8]   L. J. R. Nunes, M. Casau, M. F. Dias, J. C. O. Matias, and L. C. Teixeira, "Agroforest woody residual biomass-to-energy supply chain analysis: Feasible and sustainable renewable resource exploitation for an alternative to fossil fuels," Results Eng., vol. 17, p. 101010, Mar. 2023, doi: 10.1016/j.rineng.2023.101010.

[9]   D. Rajendran, P. B. Ezhuthupurakkal, R. Lakshman, N. K. S. Gowda, A. Manimaran, and S. B. Rao, "Application of encapsulated nano materials as feed additive in livestock and poultry: a review," Vet. Res. Commun., vol. 46, no. 2, pp. 315–328, Jun. 2022, doi: 10.1007/s11259-022-09895-7.

[10]   A. Ferreira et al., "Brazilian poultry activity waste: Destinations and energetic potential," Renew. Sustain. Energy Rev., vol. 81, pp. 3081–3089, Jan. 2018, doi: 10.1016/j.rser.2017.08.078.

[11]   A. R. K. Gollakota, N. Kishore, and S. Gu, "A review on hydrothermal liquefaction of biomass," Renew. Sustain. Energy Rev., vol. 81, pp. 1378–1392, Jan. 2018, doi: 10.1016/j.rser.2017.05.178.

[12]   J. Ni et al., "A review on fast hydrothermal liquefaction of biomass," Fuel, vol. 327, p. 125135, Nov. 2022, doi: 10.1016/j.fuel.2022.125135.

[13]   L. B. Silva Thomsen, P. N. Carvalho, J. S. Dos Passos, K. Anastasakis, K. Bester, and P. Biller, "Hydrothermal liquefaction of sewage sludge; energy considerations and fate of micropollutants during pilot scale processing," Water Res., vol. 183, p. 116101, Sep. 2020, doi: 10.1016/j.watres.2020.116101.

[14]   L. B. Silva Thomsen, K. Anastasakis, and P. Biller, "Hydrothermal liquefaction potential of wastewater treatment sludges: Effect of wastewater treatment plant and sludge nature on products distribution," Fuel, vol. 355, p. 129525, Jan. 2024, doi: 10.1016/j.fuel.2023.129525.

[15]   Y. Ouyang, C. Yu, G. Yan, and J. Chen, "Machine learning approach for the prediction and optimization of thermal transport properties," Front. Phys., vol. 16, no. 4, p. 43200, Aug. 2021, doi: 10.1007/s11467-020-1041-x.

[16]   H. Zhong et al., "CFD-based reduced-order modeling of fluidized-bed biomass fast pyrolysis using artificial neural network," Renew. Energy, vol. 152, pp. 613–626, Jun. 2020, doi: 10.1016/j.renene.2020.01.057.

[17]   B. V. Ayodele, S. I. Mustapa, R. Kanthasamy, M. Zwawi, and C. K. Cheng, "Modeling the prediction of hydrogen production by co-gasification of plastic and rubber wastes using machine learning algorithms," Int. J. Energy Res., vol. 45, no. 6, pp. 9580–9594, May 2021, doi: 10.1002/er.6483.

[18]   T. Yang et al., "An outbreak of Getah virus infection among pigs in China, 2017," Trans bound. Emerg. Dis., vol. 65, no. 3, pp. 632–637, Jun. 2018, doi: 10.1111/tbed.12867.

[19].   Oliveira, A.L., et al., GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. Information and Software Technology, 2010. 52(11): p. 1155-1166.

[20].   Heimann, K. and R. Huerlimann, Microalgal classification: major classes and genera of commercial microalgal species, in Handbook of marine microalgae. 2015, Elsevier. p. 25-41.

[21].   Marks-Bielska, R., et al., The importance of renewable energy sources in Poland's energy mix. Energies, 2020. 13(18): p. 4624.

[22].   Mohammadi, A. and M. Mehrpooya, A comprehensive review on coupling different types of electrolyzer to renewable energy sources. Energy, 2018. 158: p. 632-655.

[23].   Abdulyekeen, K.A., et al., Torrefaction of biomass: Production of enhanced solid biofuel from municipal solid waste and other types of biomass. Renewable and Sustainable Energy Reviews, 2021. 150: p. 111436.

[24]. Aravind, S., et al., Conversion of green algal biomass into bioenergy by pyrolysis.A review. Environmental Chemistry Letters, 2020. 18(3): p. 829-849.

[25]. Otálora, P., et al., Microalgae classification based on machine learning techniques.Algal Research, 2021. 55: p. 102256.

[26]. Hernández-Pérez, A. and J.I. Labbé, Microalgas, cultivo y beneficios. Revista de biología marina y oceanografía, 2014. 49(2): p. 157-173.

[27]. González-López, C.V., et al., Valorization of Microalgae and Energy Resources, in Sustainable Mobility. 2020, IntechOpen.

[28]. Peng, L., et al., Biofuel production from microalgae: a review. Environmental Chemistry Letters, 2020. 18(2): p. 285-297.

[29]. Qureshi, K.M., et al., A technical review on semi-continuous and continuous pyrolysis process of biomass to bio-oil. Journal of Analytical and Applied Pyrolysis, 2018. 131: p. 52-75.

[30]. Yang, C., et al., Pyrolysis of microalgae: A critical review. Fuel Processing Technology, 2019. 186: p. 53-72.

[31]. Yang, X., et al., Simulation model of pyrolysis biofuel yield based on algal components and pyrolysis kinetics. Bioenergy Research, 2014. 7(4): p. 1293-1304.

[32]. Kim, S.-S., et al., Pyrolysis of microalgae residual biomass derived from Dunaliella tertiolecta after lipid extraction and carbohydrate saccharification. Chemical Engineering Journal, 2015. 263: p. 194-199.

[33]. Muradov, N., et al., Pyrolysis of fast-growing aquatic biomass–Lemna minor (duckweed): Characterization of pyrolysis products. Bioresource technology, 2010. 101(21): p. 8424-8428.

[34]. Perazzini, H., et al., Thermal treatment of solid wastes using drying technologies: A review. Drying technology, 2016. 34(1): p. 39-52.

[35]. Ullah, Z., et al., A comparative study of machine learning methods for bio-oil yield prediction–A genetic algorithm-based features selection. Bioresource Technology, 2021. 335: p. 125292.

[36]. Shafizadeh, A., et al., Machine learning predicts and optimizes hydrothermal liquefaction of biomass. Chemical Engineering Journal, 2022. 445: p. 136579.

[37]. Tang, Q., et al., Prediction of bio-oil yield and hydrogen contents based on machine learning method: effect of biomass compositions and pyrolysis conditions. Energy & Fuels, 2020. 34(9): p. 11050-11060.

[38]. Zhang, T., et al., Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions. Fuel, 2022. 312: p. 122812.

[39]. Zhang, W., et al., Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. Bioresource Technology, 2021. 342: p. 126011.

[40]. Quaranta, G., W. Lacarbonara, and S.F. Masri, A review on computational intelligence for identification of nonlinear dynamical systems. Nonlinear Dynamics, 2020. 99(2): p. 1709-1761.

[41]. Mutlu, A.Y. and O. Yucel, An artificial intelligence based approach to predicting syngas composition for downdraft biomass gasification. Energy, 2018. 165: p. 895-901.

[42]. Khan, M., et al., Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. Bioresource technology, 2022. 355: p. 127215.

[43]. Ray, S. A quick review of machine learning algorithms. in 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). 2019. IEEE.

[44]. Uddin, S., et al., Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 2019. 19(1): p. 1-16.

[45]. Che, D., et al., Decision tree and ensemble learning algorithms with their applications in bioinformatics. Software tools and algorithms for biological systems, 2011: p. 191-199.

[46]. Mahesh, B., Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 2020. 9: p. 381-386.

[47]. Abd El Aziz, M., et al. Prediction of biochar yield using adaptive neuro-fuzzy inference system with particle swarm optimization. in 2017 IEEE PES PowerAfrica. 2017. IEEE.

[48]. Lewis, D.P., T. Jebara, and W.S. Noble, Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics, 2006. 22(22): p. 2753-2760.

[49]. Jeng, J.-T., Hybrid approach of selecting hyperparameters of support vector machine for regression. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2006. 36(3): p. 699-709.

[50]. Dong, X., et al., A survey on ensemble learning. Frontiers of Computer Science, 2020. 14(2): p. 241-258.

[51]. Li, H., et al., A Fruit Tree Disease Diagnosis Model Based on Stacking Ensemble Learning. Complexity, 2021. 2021.

[52]. Sarker, I.H., Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2021. 2(3): p. 1-21.

[53]. Kocijan, J., K. Ažman, and A. Grancharova. The concept for Gaussian process model based system identification toolbox. in Proceedings of the 2007 international conference on Computer systems and technologies. 2007.

[54]. Schulz, E., M. Speekenbrink, and A. Krause, A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology, 2018. 85: p. 1-16.

[55]. Ngiam, K.Y. and W. Khor, Big data and machine learning algorithms for health- care delivery. The Lancet Oncology, 2019. 20(5): p. e262-e273.

[56]. Prieto, A., et al., Neural networks: An overview of early research, current frameworks and new challenges. Neurocomputing, 2016. 214: p. 242-268.

[57]. Sunphorka, S., B. Chalermsinsuwan, and P. Piumsomboon, Application of artificial neural network for kinetic parameters prediction of biomass oxidation from biomass properties. Journal of the Energy Institute, 2017. 90(1): p. 51-61.

[58]. Siregar, S.P. and A. Wanto, Analysis of artificial neural network accuracy using backpropagation algorithm in predicting process (forecasting). IJISTECH (International Journal of Information System and Technology), 2017. 1(1): p. 34-42.

[59]. Roberge, V., M. Tarbouchi, and G. Labonté, Comparison of parallel genetic algorithm and particle swarm optimization for real-time UAV path planning. IEEE Transactions on industrial informatics, 2012. 9(1): p. 132-141.

[60]. Tabassum, M. and K. Mathew, A genetic algorithm analysis towards optimization solutions. International Journal of Digital Information and Wireless Communications (IJDIWC), 2014. 4(1): p. 124-142.

[61]. Kumar, S., S. Jain, and H. Sharma, Genetic algorithms, in Advances in swarm intelligence for optimizing problems in computer science. 2018, Chapman and Hall/CRC. p. 27-52.

[62]. Albadr, M.A., et al., Genetic algorithm based on natural selection theory for optimization problems. Symmetry, 2020. 12(11): p. 1758.

[63]. García-Martínez, C., et al., Global and local real-coded genetic algorithms based on parent-centric crossover operators. European journal of operational research, 2008.