Real-Time Tracking of Renewable Carbon Content with AI-aided Approaches During Co-Processing of Biofeedstocks

Liang Cao,[†,‖] Jianping Su,[‡,‖] Jack Saddler,[‡] Yankai Cao,[†] Yixiu Wang,[†] Gary Lee,[¶] Lim C. Siang,[§] Robert Pinchuk,[¶] Jin Li,[§] and R.Bhushan Gopaluni[*,†]

†*Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada*

‡*Forest Products Biotechnology/Bioenergy Group, The University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada*

¶*Department of Low Carbon Strategy, Burnaby Refinery, BC V5C 1L7, Canada*

§*Department of Process Control Engineering, Burnaby Refinery, BC V5C 1L7, Canada*

‖*Equal contribution author.*

E-mail: bhushan.gopaluni@ubc.ca

Abstract

Decarbonization of the oil refining industry is essential for reducing carbon emissions and mitigating climate change. Co-processing bio feed at existing oil refineries is a promising strategy for achieving this goal. However, accurately quantifying the renewable carbon content of co-processed fuels can be challenging due to the complex process involved. Currently, it can only be achieved through expensive offline ^{14}C measurements. To address this issue, with high-quality and large-scale commercial data, our study proposes a novel approach that utilizes data-driven methods to build inferential

sensors, which can estimate the real-time renewable content of biofuel products. We have collected over 1,000,000 co-processing data points from refineries under different bio feed co-processing ratios and operational conditions—the largest dataset of its kind to our knowledge. We use interpretable deep neural networks to select model inputs, then apply robust linear regression and bootstrapping techniques to estimate renewable content and confidence interval. Our method has been validated with four previous ^{14}C measurements during co-processing at the fluid catalytic cracker. This novel methods provides a practical solution for the industry and policymakers to quantify renewable carbon content and accelerate the transition to a more sustainable energy system.

1. Introduction

Our society's dependence on fossil fuels is one of the main causes of climate change, and its continued use has led to many problems, such as increased extreme weather events, ecological damage, forest fires, and glacier melting. According to the International Energy Agency (IEA), in 2020, fossil fuels accounted for approximately 80% of the world's primary energy consumption.[1] Achieving the net-zero goal and decarbonizing the economy requires a fundamental shift from a fossil-fuel-based economy to a renewable energy-dominated economy.[2]

The current pace of decarbonization by building from scratch is accelerating (such as solar photovoltaic system, wind farms, and electric cars), but still much slower than expected. Thus, to further accelerate the process in a cost effective way, working with the current fossil fuel industry is a quicker and reliable path forward. Co-processing is a technology that enables oil refineries to utilize into biorefineries by co-processing low-carbon-intensive feedstocks, such as bio-crudes made from forest, mill residues,[3,4] microalgae, municipal sludge, and municipal waste.[5] These materials can undergo processes such as hydrothermal liquefaction to yield biocrude, which is then suitable for further processing.[6,7] It has a lower carbon footprint than conventional petroleum refining since burning bio feed-derived fuel is carbon neutral (although producing the biofuel will have some greenhouse gas emissions).

Co-processing oleochemical feedstocks have been fully commercialized by various refiners around the world, either in the hydrotreater[8] or at the fluid catalytic cracker.[9] This provides valuable experiences for refiners to co-process biocrudes made from waste feedstocks in the future. As the world demands lower carbon intensity fuels, the renewable content of these fuels will become another essential property, much like sulfur, nitrogen, and flash point are for the fuels produced today. This means that fuel producers will need to consider the renewable content of their products as a critical factor. Furthermore, policies like the Low Carbon Fuel Standard (LCFS) incentivize refiners to co-process renewable feedstocks with conventional crude oil.[10] To ensure the effectiveness of policies like LCFS in influencing the oil sector, it is crucial to estimate accurately the reduction in carbon intensity achieved through co-processing renewable feedstocks with conventional crude oil.

The task of accurately tracking the renewable content of fuels resulting from co-processing of biogenic feedstocks with fossil fuels poses significant challenges.[3,11] A key issue in this regard is the uneven distribution of green molecules across different fractions combined with the limited availability of effective techniques for quantifying the renewable content of each stream. One common approach for tracking the renewable content from fossil molecules is the isotopic analysis, specifically radiocarbon dating using the ^{14}C isotope.[9,12] This technique, widely used in the field of archaeology to determine age, can be adapted to trace the origin of carbon molecules. Biofuels were produced in recent years while fossil fuels were formed hundreds of millions of years ago. This property makes ^{14}C a useful tool for quantifying the renewable content of biofuels. However, if multiple measurements are needed, the equipment setup and maintenance costs can be very high (thus only a few dedicated labs can perform such tests). Additionally, this method can not provide continuous measurements which means the test results only provide a snapshot of the operation. Another potential solution is the development of online renewable content monitoring systems to provide continuous measurements. However, the implementation of such equipment requires extensive evaluation before oil refiners can fully adopt them.

The co-processing process is highly complex, involving multiple interrelated unit operations and control loops. The process exhibits high-dimensional, nonlinear, and dynamic characteristics. The emergence of artificial intelligence technology has brought new ideas and solutions for the monitoring and optimization of this type of industrial processes.[13–15] Artificial intelligence technology can learn and analyze a large amount of historical data to establish accurate soft sensor models, thereby achieving real-time monitoring and prediction of process status andperformance.[16–18]

The soft sensor is a data-driven artificial intelligence method designed to predict and monitor key performance indicators in the production process.[17,18] Soft sensors are usually built using machine learning algorithms such as artificial neural networks, support vector machines, and decision trees. They have become an essential tool for process monitoring and control, leading to improved efficiency, product quality, and reduced downtime. It takes existing external variables as inputs and builds mathematical models to predict and estimate the values of some key process parameters or variables, thus enabling real-time monitoring and control.

Our industrial partner, Parkland Refining Ltd, is currently co-processing oleochemical/lipid feedstocks such as tallow, canola oil, and tall oil that reduce the carbon intensity (CI) of the various fuels that they produce. This is similar to what other oil companies around the world, such as BP in Washington State, Preem in Sweden, and ENI in Italy are doing to reduce their carbon footprint. Parkland is expanding its co-processing to 5,500 barrels per day.[19] The availability of a significant amount of industrial data generated through commercial operations provides unique opportunity to build more reliable and robust AI models that can effectively track the renewable content during FCC co-processing.

In the sections that follow, we'll detail the innovations in our study. These aspects differentiate our work from typical research in the field, provide valuable contributions and new perspectives to the existing research.

For the theoretical innovations, this study is pioneering in its domain, emphasizing the

critical role of artificial intelligence in decarbonizing the oil refining industry. This study proposes an approach that can contribute to the innovative fusion of interpretable deep neural networks, robust linear regression and bootstrapping techniques to achieve high accuracy in estimating the real-time renewable content of liquids produced, signifying a leap towards the continuous, real-time monitoring of renewable carbon. The efficacy is further validated with four ^{14}C measurements employed during co-processing at the FCC. With an average error rate below 4%, we have demonstrated that AI can play a significant role in enhancing the precision of renewable carbon tracking.

For dataset contribution, we have amassed over 1,000,000 co-processing data points from refineries. This dataset, to our best knowledge, is the largest of its kind. While traditional laboratory-scale studies are often limited to a few or tens of samples, our dataset stands out in its depth and breadth. By encapsulating varied operational conditions such as temperature and catalyst activity, along with diverse bio-oil inclusion ratios, our data is pivotal in driving forward significant advancements in the domain of bio feed co-processing.

In practical implications, our findings provide oil refineries with an invaluable toolset to quantify renewable carbon accurately. This can help policymakers and stakeholders better understand the performance of bio feed in renewable energy production, informing decisions on promotion and investment in the industry. Furthermore, our method offers substantial cost reductions for refineries (millions of dollars) and ensures companies secure optimal governmental incentives through accurate renewable carbon measurements. This precision not only provides financial benefits but also propels the broader shift towards renewable energy sources.

2. Co-processing at the FCC

Co-processing is a process that involves the simultaneous treatment of liquefied biomass and petroleum-based feedstocks (such as crude oil or natural gas liquids) in a refinery. In this

process,the liquefied biomass feedstock is derived from the bioprocessing of solid biomass materials, such as agricultural waste, forestry residues, crop residues, and organic waste, to yield compatible intermediates suitable for refinery operations. These intermediates are then treated to meet essential contaminants and flow specifications before being mixed with petroleum-based feedstocks. The resulting mixture is then subjected to various refining processes, such as hydrotreating and catalytic cracking, to produce biofuels (such as renewable diesel, gasoline, and jet fuel) and other value-added products.

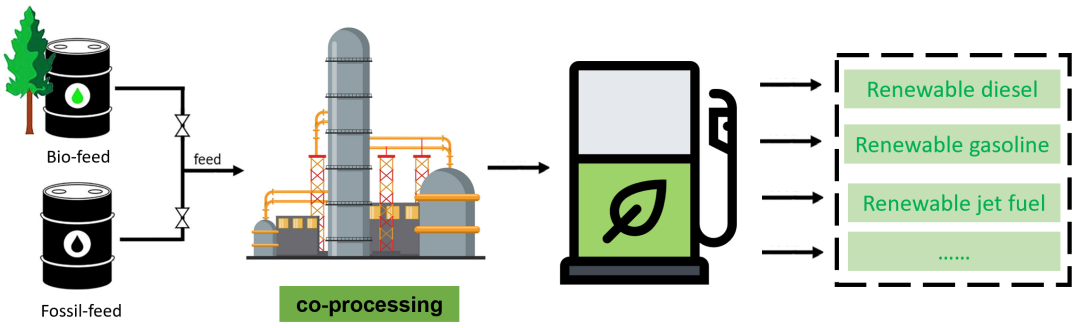

Figure 1: A diagram of co-processing

Co-processing has numerous advantages, including reducing dependence on oil, lowering carbon emissions, promoting sustainable development of agriculture and forestry, and the diversification of energy sources. The technical and economic potential of co-processing is also constantly improving. With the advancement of technology and the realization of economies of scale, it will gradually become an essential method of energy and chemical production.

Fluid catalytic cracker (FCC) is a core process in a refinery. It consists of three main parts, namely the reactor, the regenerator, and the fractionator, which can be seen in Figure 2. It is an intermediate unit that processes the heavy hydrocarbons from crude oil and "cracks" them into smaller hydrocarbons, which can then be processed into a wide variety of different products,[9]including light catalytic cracking oil (LCC), heavy catalytic cracking oil (HCC), liquefied petroleum gas (LPG), light cycle oil (LCO), heavy cycle oil (HCO) and decant oil.

It is highly efficient and cost-effective, and can be used to process a wide range of feedstocks.

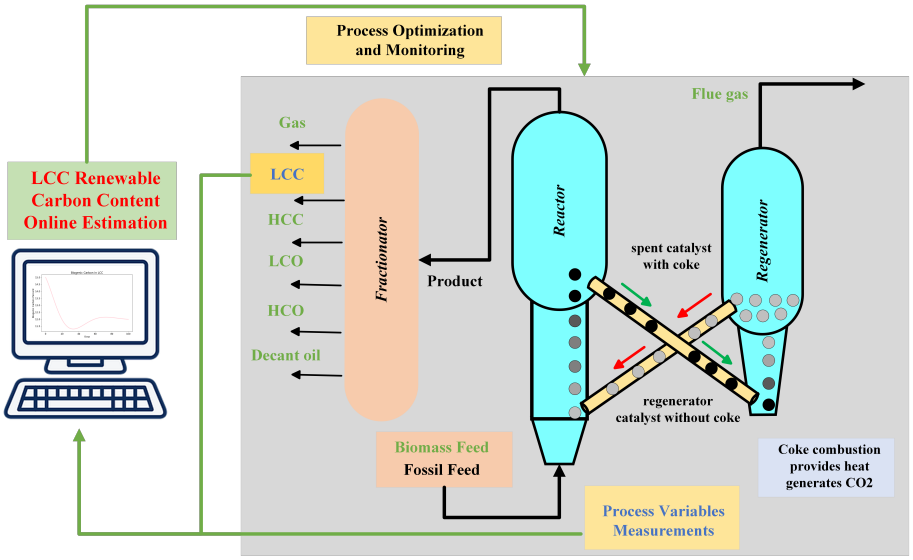

Figure 2: A flow diagram of a Fluid Catalytic Cracking unit

Biogenic feedstocks can be inserted either at the hydrotreater or the FCC. Co-processing at the FCC has been commercialized using lipid feedstocks, as the FCC provides greater flexibility and can process much higher amounts of biogenic feedstocks compared to the hydrotreater without requiring significant upgrades. Moreover, the FCC appears to be a more viable insertion point for future biocrudes or even pyrolyzed plastic waste, potentially leading to the development of green plastics for integrated refineries. In summary, the FCC offers greater flexibility with respect to the quality of crude and may be a more suitable option for co-processing certain feedstocks compared to hydrotreating.

3. AI for Renewable Carbon Tracking

The integration of Artificial Intelligence (AI) into the field of bio feed co-processing has the potential to provide new and innovative solutions for tracking the renewable carbon. In this work, we will focus on the application of soft sensors for modeling and predicting renewable carbon in real-time. By utilizing deep learning for feature selection and combining bootstrapping and robust linear regression for building robust soft sensor models, we aim to

use AI to track renewable carbon content in bio feed co-processing.

3.1 Introduction to Soft Sensors

Soft sensors are software-based sensors designed to estimate specific quality variables that are difficult or expensive to measure in real time, like product composition, and distillation temperature. These variables are usually very important since they are highly related to product quality and process safety. If one can't measure those variables, it is challenging to control and monitor their quality. Figure 3 gives a diagram of soft sensor.

The basic idea of soft sensor is to select easily measurable variables X to construct a mathematical relationship that can estimate the values of quality variables y. The mathematical definition of a soft sensor can be given as follows:

$$y = f\left(d, X\right) \tag{1}$$

where d, X, and y are unknown noise, easy-to-measure variables, and quality variables, respectively.

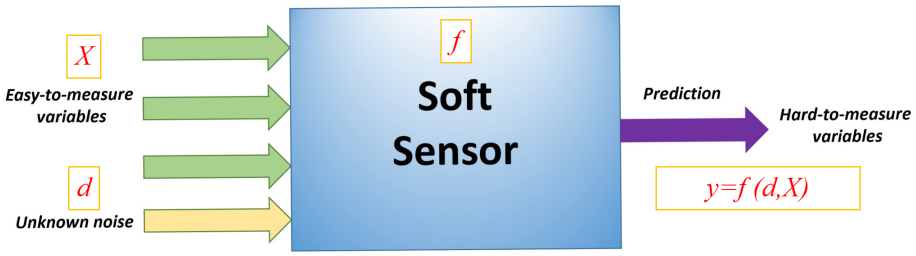

Figure 3: A diagram of soft sensor

As shown in Equation 1, we can see the use of soft sensors in tracking renewable carbon involves two main components: the selection of easy-to-measure process variables X and the establishment of regression models f. The selection of process variables is vital for the soft sensor's accuracy, while the construction of regression models enables real-time predictions of renewable carbon. In subsequent sections, we will examine each component in detail and

focus on the methods employed in tracking renewable carbon.

3.2 Soft Sensors: Process Variables Selection

In process variable selection, including all variables not only makes the soft sensor complex, but also leads to over-fitting, which reduces generalization ability and affects the accuracy of online prediction. A common practice is to select variables that are highly correlated with quality variables, using methods such as correlation and mutual information.[20] However, co-processing is characterized by strong and unknown nonlinearities as well as complex dynamic processes, which cause most traditional process variable selection algorithms to yield unsatisfactory performance. In contrast, deep neural network (DNN) offers robust nonlinear fitting and dynamic modeling capabilities, making it a suitable alternative.[14,15] It is a sub-field of machine learning inspired by studying brain structure and its function, and has been widely studied in process modeling.

3.2.1 Deep Neural Networks

Deep Neural Networks (DNNs) are structured with a multitude of interconnected nodes, commonly referred to as neurons. The input layer receives the data, while the output layer is responsible for generating predictions. The layers between the input and output are known as hidden layers. Each neuron within this network takes in inputs from its predecessor neurons, computes a weighted sum of these inputs, and then applies a distinct non-linear activation function to yield its output. Formally, the output for any neuron can be described as:

$$a_j^{(l)} = f_1\left(\sum_{i=1}^{n^{(l-1)}} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right) \tag{2}$$

In this equation, $a_j^{(l)}$ denotes the output of the j-th neuron in the l-th layer, $w_{ij}^{(l)}$ signifies the weight connecting the i-th neuron in the $(l-1)$-th layer to the j-th neuron in the l-th layer, $a_i^{(l-1)}$ stands for the output of the i-th neuron in the $(l-1)$-th layer, $b_j^{(l)}$ is the bias

term associated with the j-th neuron in the l-th layer, $n^{(l-1)}$ indicates the total neurons in the $(l-1)$-th layer, and f_1 is the designated activation function.

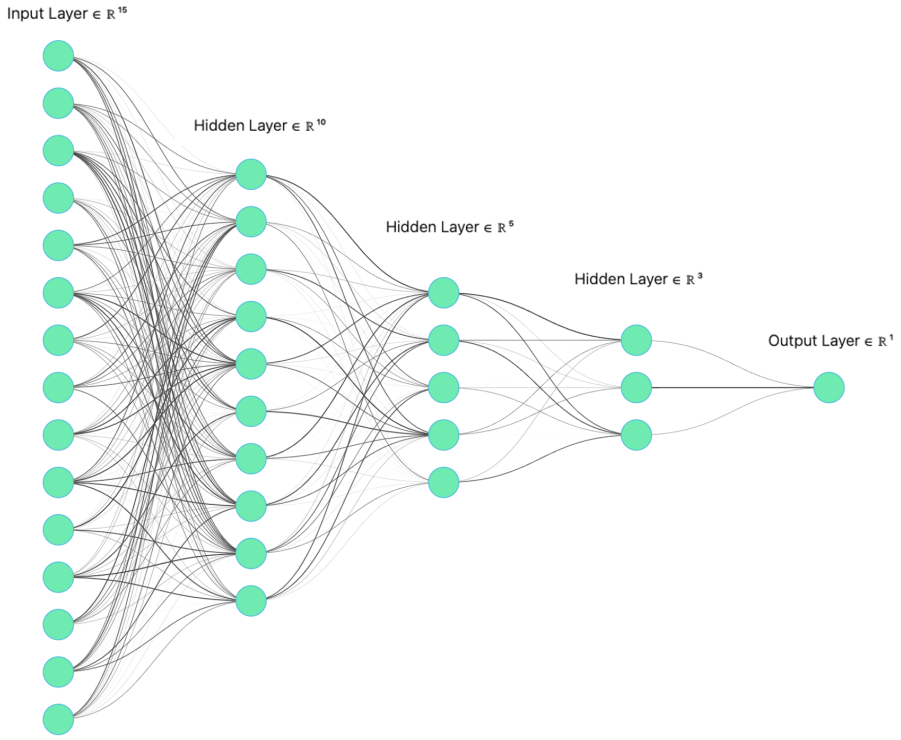

Figure 4: A schematic representation of DNN structure

Figure 4 depicts the DNN structure employed in our study. The input layer consists of 15 neurons, followed by three hidden layers having 10, 5, and 3 neurons respectively. The architecture is finalized with an output layer containing a singular neuron. In this configuration, the hyperbolic tangent is the chosen activation function. Optimization is governed by the Adam algorithm, with the primary objective being the reduction in the mean squared error between predicted and actual LCC.

One challenge with DNNs is their inherent complexity, as highlighted by our model which has 237 parameters for this simple modeling problem. Moreover, they often lack clarity, making it challenging to determine the importance of specific features and the logic behind predictions.[21,22] To address these challenges, we turn our attention to SHAP (SHapley Additive exPlanations), drawing inspiration from cooperative game theory.[23] By integrating SHAP into our analysis framework, we unpack the black-box nature of DNNs and has a

better understanding of how the DNN is making its predictions and identifying important features.[24]

3.2.2 SHAP (SHapley Additive exPlanations)

SHAP is a technique used to interpret predictions made by machine learning models. It offers a consistent method for assessing the impact of each feature on a prediction, factoring in both the feature's value and its interactions with other features. SHAP values are derived from the concept of Shapley values in cooperative game theory, which ensures a fair distribution of a coalition's contribution among its members. In the realm of machine learning, the coalition refers to the collection of features, and the contribution corresponds to the model's prediction. The Shapley value of feature x is defined as follows:

$$\phi_x\left(f_3\right)=\sum_{S^{\prime}\subseteq p\backslash x} w_x\left(S^{\prime}\right)\left[f_3\left(S^{\prime}\cup\{x\}\right)-f_3\left(S^{\prime}\right)\right] \tag{3}$$

where f_3 is the complex model like DNNs, $\phi_x(\bullet)$ is the Shapley value of feature x under model f_3, p is the number of input features, $S^{\prime}$ is a subset of the features. $S^{\prime}\cup x$ represents the union of the subset $S^{\prime}$ and the single-element set containing the feature x. In other words, it combines the features in the subset $S^{\prime}$ with the feature x to form a new set that includes all the elements from both $S^{\prime}$ and x. $w_x\left(S^{\prime}\right)$ is defined as $\frac{|S^{\prime}|!(p-|S^{\prime}|-1)!}{p!}$, where $|S^{\prime}|$ represents the number of elements in the subset $S^{\prime}$. The denominator $p!$ represents all possible feature combinations; the numerator $|S^{\prime}|!(p-|S^{\prime}|-1)!$ means the appearance times of $S^{\prime}\cup x$ in all $p!$ combinations; $f_3\left(S^{\prime}\cup x\right)-f_3\left(S^{\prime}\right)$ indicates the expected marginal contribution of feature x in one combination.

Figure 5 illustrates the algorithm's flowchart for feature selection using SHAP and DNNs. By combining the strengths of DNNs and SHAP, the step of using DNNs and SHAP to select important features is given as follows:

- Train a DNN model: Train a DNN model on the data using the full set of features.

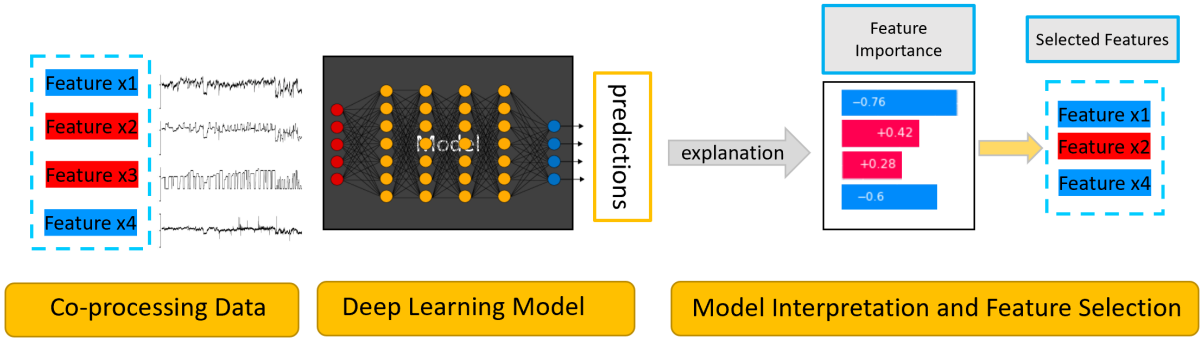

Figure 5: A diagram of feature selection with SHAP and DNNS

This model will be used to make accurate predictions.

- Use SHAP values to identify important features: SHAP values can be used to calculate the contribution of each feature to the model's prediction for each sample. The magnitude of the SHAP value indicates the importance of each feature.

- Select the most important features: This process can be guided by using a threshold of the absolute SHAP values. Specifically, given a matrix M, the objective is to ascertain the minimal subset of features whose absolute SHAP values make up more than a specified percentage (e.g., 90%) of the entire matrix's absolute SHAP values.

For achieving this, we start by defining a vector V_j for each feature j where $V_{j,i}$ represents the SHAP value of the j^{th} feature for the i^{th} sample. The importance measure of each feature is then the summation of the absolute values of its SHAP values: $S_j = \sum_{i=1}^{n} |V_{j,i}|$, with n denoting the number of samples. Subsequently, features are sorted based on S_j, producing an index list L with the order $S_{L_1} \geq S_{L_2} \geq ... \geq S_{L_m}$, where m symbolizes the number of features. Our goal becomes to identify the minimal k such that the initial k features in list L have their absolute SHAP values contributing up to predetermined threshold T of the total absolute SHAP values. Mathematically, this can be expressed as:

$$\frac{\sum_{i=1}^{k} S_{L_i}}{\sum_{i=1}^{m} S_{L_i}} \geq p \tag{4}$$

The selected features, as identified through this method, are subsequently utilized for the Huber regression analysis detailed in the following section.

3.3 Soft Sensors: Bootstrapping and Huber Regression

In this section, we outline the use of bootstrapping[25] and Huber regression[26] to develop soft sensor models for the online prediction of renewable carbon in bio feed co-processing. Our choice for these methods is based on several considerations:

Bootstrapping Benefits: Bootstrapping uses re-sampled data, allowing us to create multiple models. This not only improves stability but also helps us quantify model uncertainty, provide confidence intervals, and offer a deeper understanding of the model predictions.

Simplicity of Huber Regression: This method is both robust and interpretable, simplifying the task of understanding relationships between variables. It directly shows how input features relate to the outcome. Although we use DNNs for feature selection, they can be more complex and less interpretable in tracking renewable carbon during co-processing. We prefer simple models over complicated models as they can be easily maintained within the refinery. It is also more stable than complex models.

In Huber regression, the Huber loss function is commonly employed. This function combines the squared loss (for small errors) and linear loss (for large errors) and is defined as:

$$L_{huber}(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta \cdot (|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

where $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$, δ is a tuning parameter that determines the threshold at which the linear loss function is used. The Huber loss function assigns less weight to outliers than the squared error loss function, thus enhancing model robustness.

Building upon the robust foundation set by the Huber loss function, bootstrapping further augments this robustness. By defining M as the number of bootstrapping iterations, the final

model's coefficients can be deduced by averaging those across all bootstrapped models:

$$\beta_k = \frac{1}{M}\sum_{i=1}^{M}\beta_{k,i} \qquad k = 0, 1, 2, ..., p \tag{5}$$

Overall, the combination of Huber regression and bootstrapping techniques provides a powerful tool for the prediction of renewable carbon that can effectively deal with the complexities and uncertainties present in bio feed co-processing. Figure 6 shows the methodological steps of our algorithmic process.

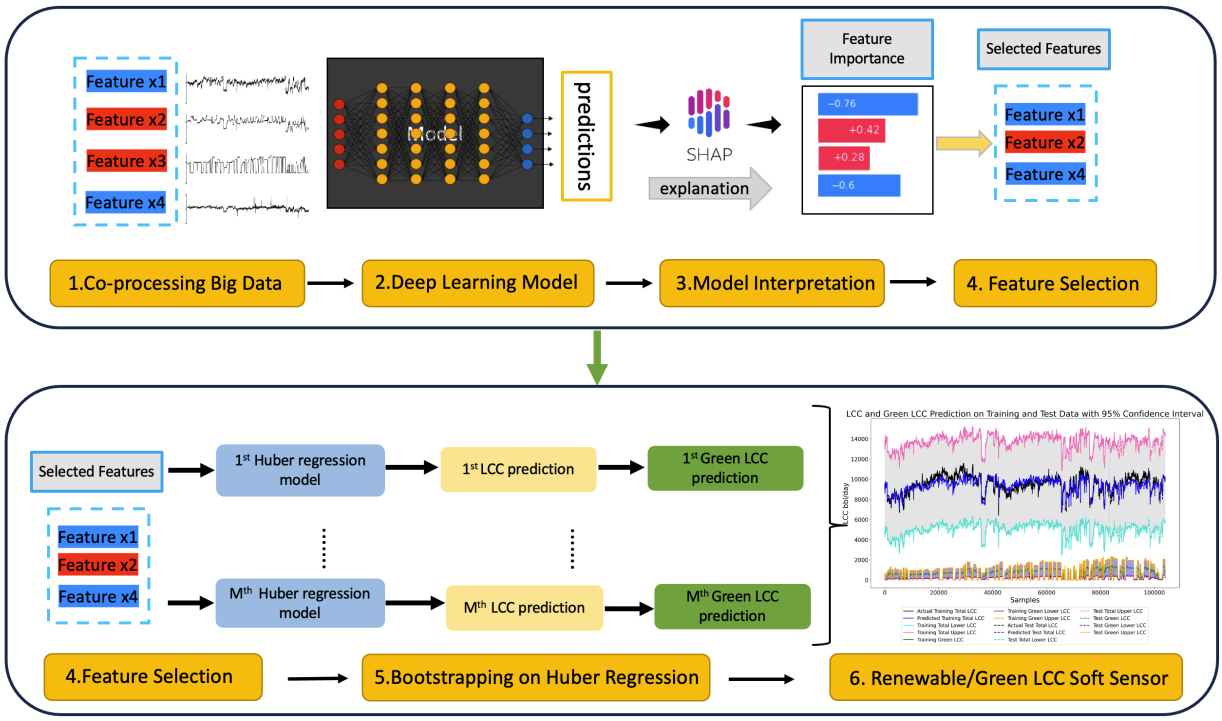

Figure 6: Process Diagram of Renewable Carbon Tracking with AI-aided Method

4. Results and discussion

4.1 Data gathering and preprocessing

In this section, we utilized commercial-scale co-processing data from the Parkland refinery in Burnaby, British Columbia, Canada, for our case study. It is important to highlight the

value and uniqueness of the data we have collected for our study. This is a commercial dataset that provides a large volume of high-quality data from continuous co-processing operations at a commercial scale. Its commercial nature ensures that the data reflects real-world scenarios and practical considerations, which make it a valuable resource for researchers and industry professionals to develop accurate models and improve the understanding of the co-processing process. All algorithms and data will be released on GitHub upon acceptance of the manuscript for journal publication.

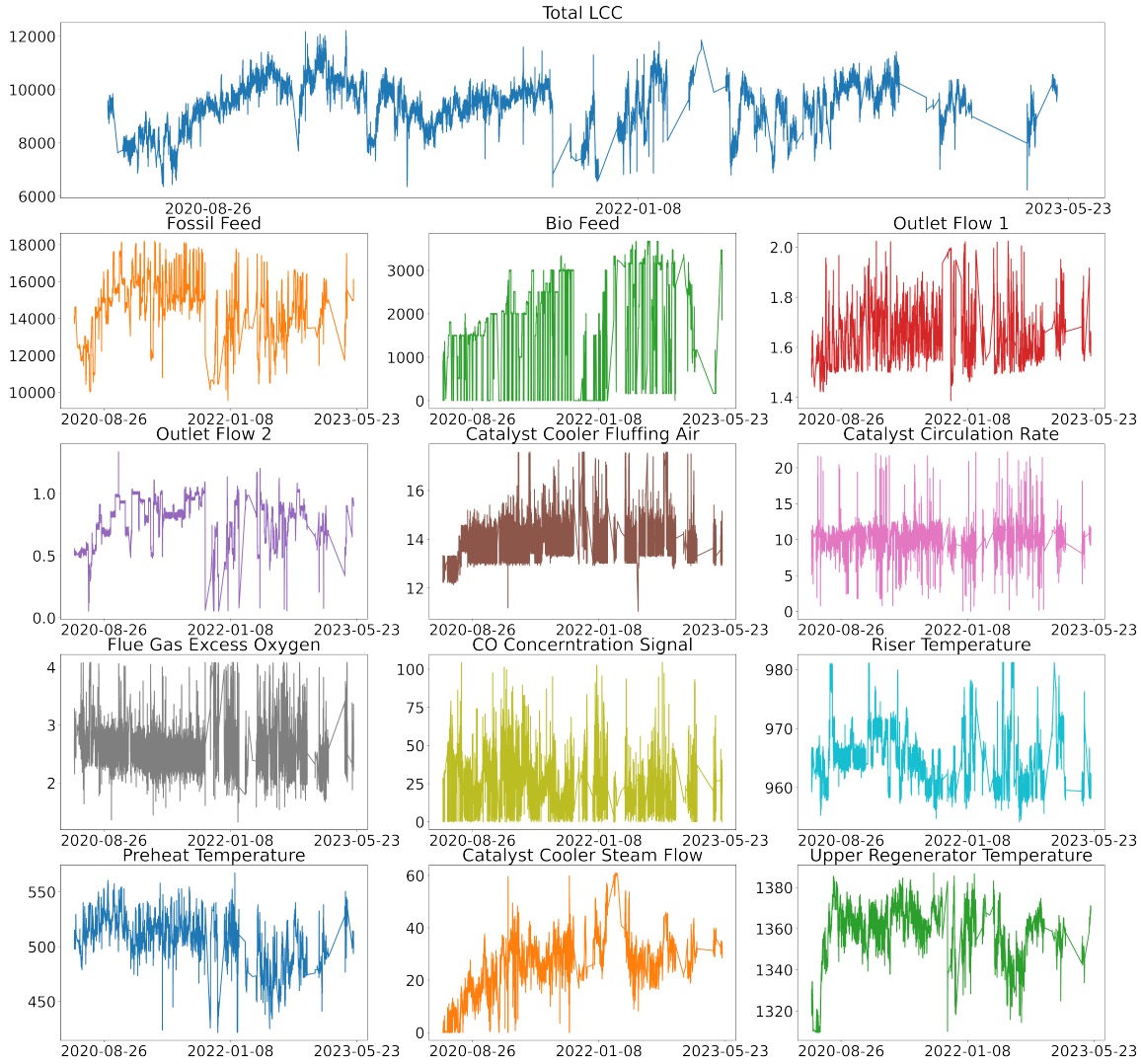

Figure 7: Visualization of co-processing data

The data from May 1, 2020 to May 1, 2023 is collected, with a total of 16 variables,

including 15 inputs and one output. The input variables consist of both fossil feed and bio feed, as well as other key variables that are particularly important in the FCC co-processing, such as the catalyst rate, reaction temperature, and others. In the data preprocessing process, we used the 3-sigma rule to remove outlier points and unreasonable data points from an engineering perspective. Ultimately, we obtained 104000 samples (80% for training and 20% for test), which will be used for subsequent analysis and modeling. Figure 7 shows the normalized process data.

The objective of this study was to track renewable carbon during the FCC bio feed co-processing. As an illustrative example, we analyze the main product of FCC, LCC, due to space constraints. LCC is a liquid petroleum product that is separated from the distillation tower after the cracking reaction occurs in the FCC unit. It is often used as a blending feedstock for gasoline. During co-processing, LCC is jointly produced by both fossil feed and bio feed. Specifically, we aimed to determine how much of the LCC production is attributable to bio feed, which is considered renewable carbon when combusted.

4.2 Process Variables Selection of LCC

In our analysis, we compared various models to identify the most appropriate variables selection methods. While various algorithms and methods were compared in our analysis, a detailed description of each method is beyond the scope of this study. Our primary focus remains on the results and their implications, as presented in Table 1. We observed that Interpretable NN is the most reasonable model, with 'Fossil Feed' and 'Bio Feed' were the two most critical features, which was consistent with real-world scenarios. On the right side of Figure 8, the specific rankings are presented. The left side of Figure 8 displays the correlations between all variables. When using correlation-based methods, bio feed was considered a relatively less important feature, incorrectly suggesting that it might not be effective in the co-processing process. This misrepresentation point that these correlations do not necessarily capture the true importance of the features.

Table 1: Feature Importance Ranking of Fossil Feed and Bio Feed in Different Models

Model	Fossil Feed	Bio Feed
Interpretable NN	**1**	**2**
LightGBM	1	6
Pearson Correlation	2	10
Mutual Information	2	3
Boruta	2	14
Random Forest	2	12
Gradient Boosting	2	6
XGBoost	2	7

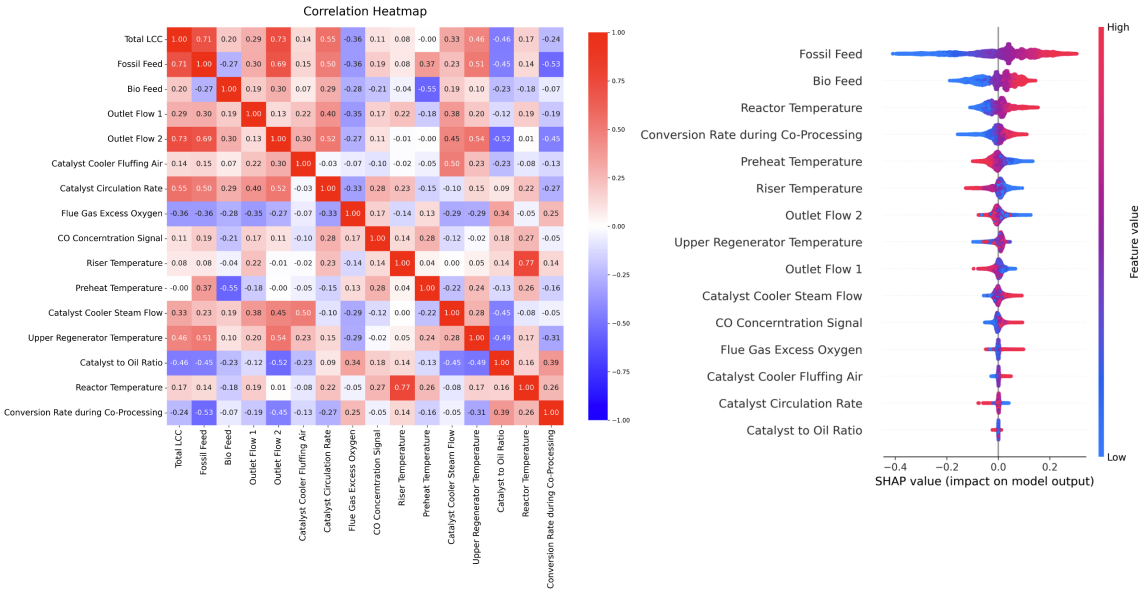

Figure 8: Process variables selection of LCC

With the feature importance ranking of Interpretable NN model, we then assessed the LCC prediction model performance by setting the predetermined threshold T as 80%. As a result, the first five features are chosen for modeling. With regard to other feature selection methods, our strategy was to include all the features deemed more critical than the 'Bio Feed'. This approach ensures that our models capture the key influencers, especially when 'Bio Feed', which aligns with real-world understanding, is not the top-ranked feature in some methods. For instance, in the LightGBM model, we incorporated the top six features, considering 'Bio Feed' was ranked sixth in terms of importance. Similarly, with the Pearson

Correlation method, where 'Bio Feed' was ranked tenth, we incorporated the top ten features for our analyses.

4.3 LCC Prediction and Renewable LCC Tracking

After data preprocessing and feature selection steps, we began predicting LCC using various machine learning regression models. We tested models like Ordinary Least Squares (OLS), Least Angle Regression (LARS), Bayesian Ridge Regression, Huber Regression, Partial Least Squares (PLS), Support Vector Regression (SVR), a three-layer neural network, and the advanced Transformer model. Each model was tested across all feature selection methods.

Based on Table 2, it's clear that most models performed better with feature selection. This improvement is seen in the lower RMSE and higher R^2 values. Among all combinations, using the Interpretable Neural Network for feature selection with the Huber Regression method showed the best results, with an RMSE of 357.05 and an R^2 value of 0.83. This means that Huber Regression, especially when combined with feature selection from the Interpretable Neural Network, is the best choice for LCC prediction in our dataset. The range of the measured output can be referenced in Figure 13, where the LCC is seen to vary predominantly between 8000 bbl/day and 11000 bbl/day. Given LCC's range, RMSE of 357.05 corresponds to approximately 3.25% to 4.46%. Figure 9 and 10 provide a visual representation of the performance of various regression models paired with different feature selection techniques, evaluated by RMSE and R^2, respectively.

Table 2: Comparison of LCC prediction with various feature selection and regression methods

	OLS		LARS		Bayesian Ridge		Huber	
Feature Selection	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
No Feature Selection	742.11	0.19	869.45	-0.11	745.89	0.19	578.21	0.52
Interpretable NN	573.62	0.52	573.62	0.52	573.62	0.52	357.05	0.83
LightGBM	569.34	0.53	569.34	0.53	569.37	0.53	472.15	0.69
XGBoost	652.12	0.38	652.12	0.38	652.29	0.38	562.91	0.54
Pearson Correlation	455.98	0.71	455.98	0.71	455.45	0.71	423.67	0.75
Mutual Information	467.92	0.69	467.92	0.69	476.27	0.68	519.48	0.61
Boruta	739.38	0.20	1847.21	-4.07	741.03	0.20	496.58	0.65
Random Forest	722.15	0.24	10255.08	-155.55	722.24	0.24	536.88	0.59
Gradient Boosting	569.43	0.53	569.43	0.53	569.48	0.53	412.12	0.76

	PLS		SVR		Neural Network		Transformer	
Feature Selection	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
No Feature Selection	755.32	0.17	567.11	0.54	628.36	0.43	858.06	-0.08
Interpretable NN	575.69	0.52	560.38	0.55	603.32	0.47	873.41	-0.12
LightGBM	551.84	0.56	565.40	0.54	460.79	0.70	873.18	-0.12
XGBoost	646.02	0.39	574.53	0.53	528.14	0.60	875.35	-0.13
Pearson Correlation	487.25	0.66	498.54	0.65	590.89	0.50	873.64	-0.12
Mutual Information	468.09	0.69	503.72	0.64	474.27	0.68	870.54	-0.11
Boruta	747.21	0.18	568.20	0.54	710.64	0.26	873.89	-0.12
Random Forest	744.00	0.19	576.01	0.52	479.48	0.67	874.59	-0.12
Gradient Boosting	579.04	0.52	573.65	0.53	584.59	0.51	866.70	-0.10

Then, we applied bootstrapping with 5000 iterations in Huber regression to estimate the uncertainty in the LCC predictions. For each iteration, a new set of model parameter estimates was generated. These results were combined to obtain the final estimates and confidence intervals according to Equation (5). This approach enabled us to establish 95% confidence intervals for the predicted LCC values. For each iteration, we assumed that LCC

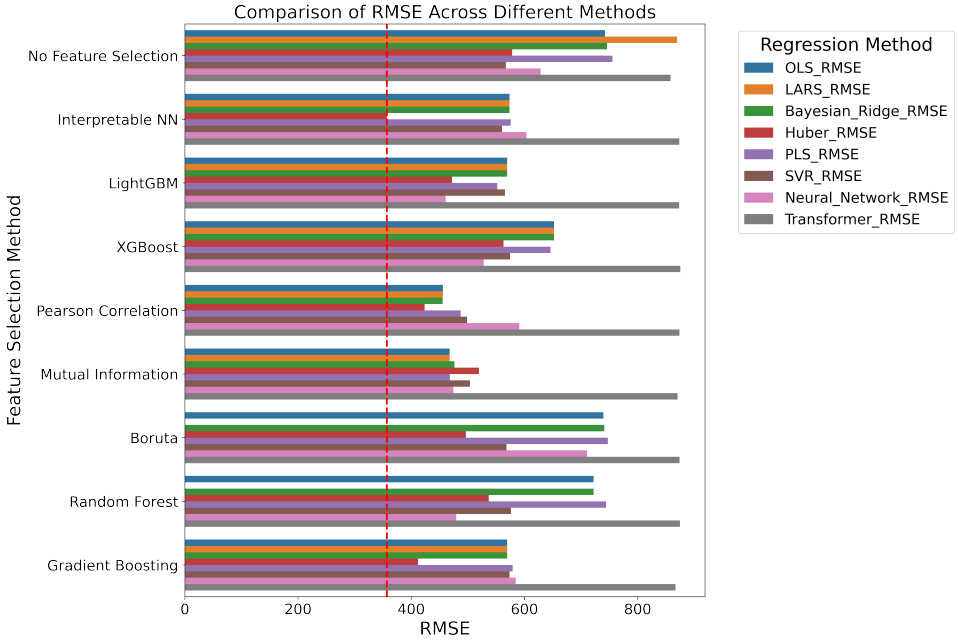

Figure 9: Comparative performance of regression models with various feature selection techniques based on RMSE

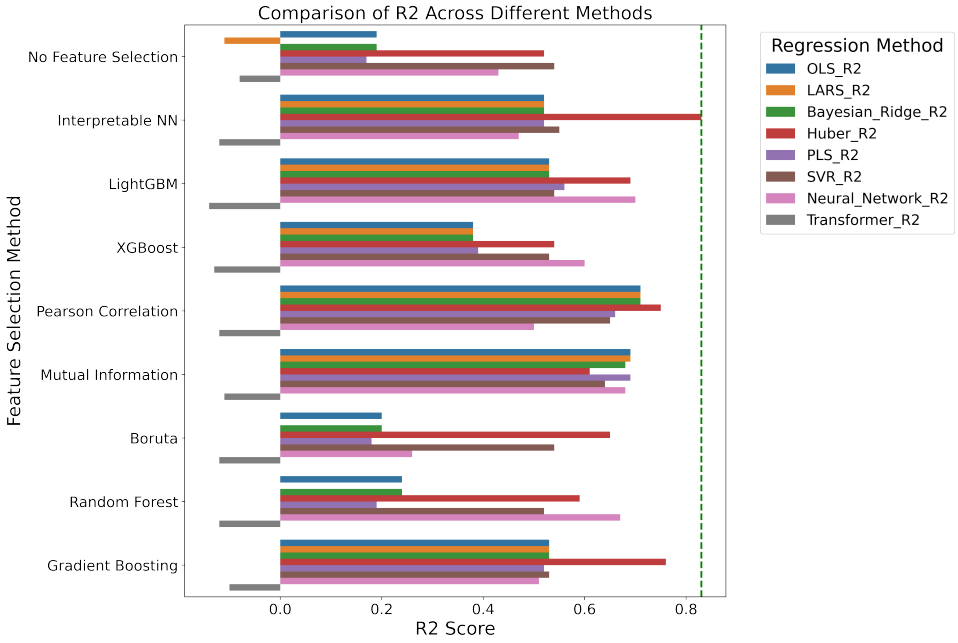

Figure 10: Comparative performance of regression models with various feature selection techniques based on R2

is a linear combination of the input variables, i.e.,

$$LCC = \underbrace{\boldsymbol{a} \cdot fossil\ feed + \boldsymbol{b} \cdot bio\ feed}_{\mathrm{LCC}(fossil,\ bio)} + \varepsilon(fossil,\ bio) \tag{6}$$

$$\begin{aligned} \varepsilon(fossil,\ bio) = {} & \boldsymbol{c} \cdot Reactor\ Temperature + \boldsymbol{d} \cdot Preheat\ Temperature \\ & + \boldsymbol{e} \cdot Conversion\ Rate\ during\ Co-Processing \end{aligned} \tag{7}$$

where ***a,b,c,d,e*** are the coefficients of Huber (robust linear) regression. The predicted LCC is further decomposed into two components, LCC $(fossil, bio)$ and $\varepsilon(fossil,\ bio)$. LCC$(fossil, bio)$ represents the main contribution of fossil feed and bio feed to the LCC production, while $\varepsilon(fossil, bio)$ accounts for the additional contribution by another 3 features (among first five features for modeling, including Reactor Temperature, Conversion Rate during Co-Processing and Preheat Temperature) that is not directly explained by the fossil feed and bio feed.

The primary objective of this section is to track renewable carbon in FCC product. Define the co-processing ratio as the ratio of bio feed to the sum of bio feed and fossil feed, and represent it as r_{co}, we can further derive the formula for renewable LCC as follows:

$$\begin{aligned} Renewable\ LCC &= \boldsymbol{b} \cdot bio\ feed + r_{co} \cdot \varepsilon(fossil,\ bio) \\ &= \boldsymbol{b} \cdot bio\ feed + bio\ feed/(bio\ feed + fossil\ feed) \cdot \varepsilon(fossil,\ bio) \end{aligned} \tag{8}$$

Figure 11 displays the distributions of the fossil feed coefficient ***a*** and the bio feed coefficient ***b*** obtained from bootstrapping. The red line represents the mean of the distributions, while the region between the green lines indicates the 95% confidence interval. The mean values of the fossil feed and bio feed coefficients are 54.2% and 32.3%, respectively. Specifically, if one unit volume of fossil feed produces 54.2% units of LCC, then one unit volume of bio feed produces approximately 32.3% units of LCC. The ratio of the bio feed coefficient to the fossil feed coefficient is a crucial factor for comparing their LCC production efficiency.

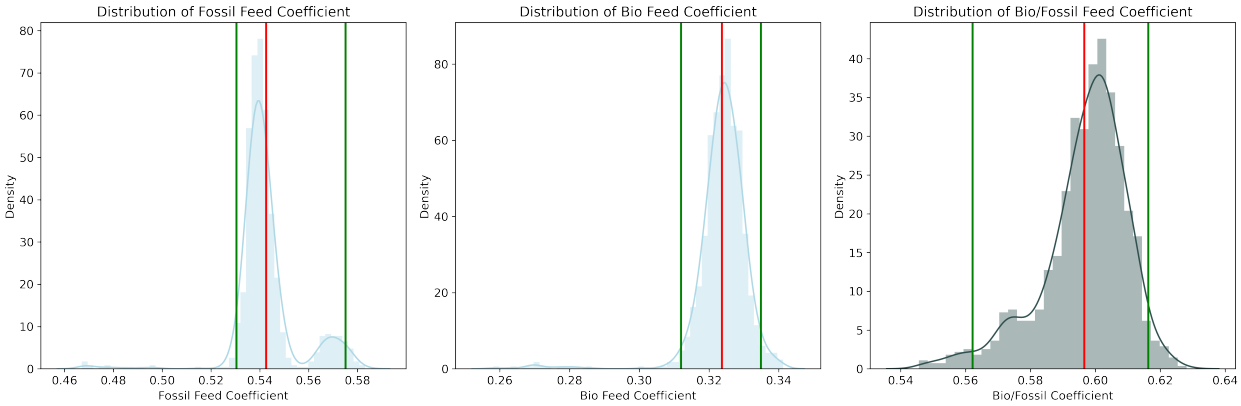

Figure 11: Distribution of fossil feed coefficient and bio feed coefficient in bootstrapping

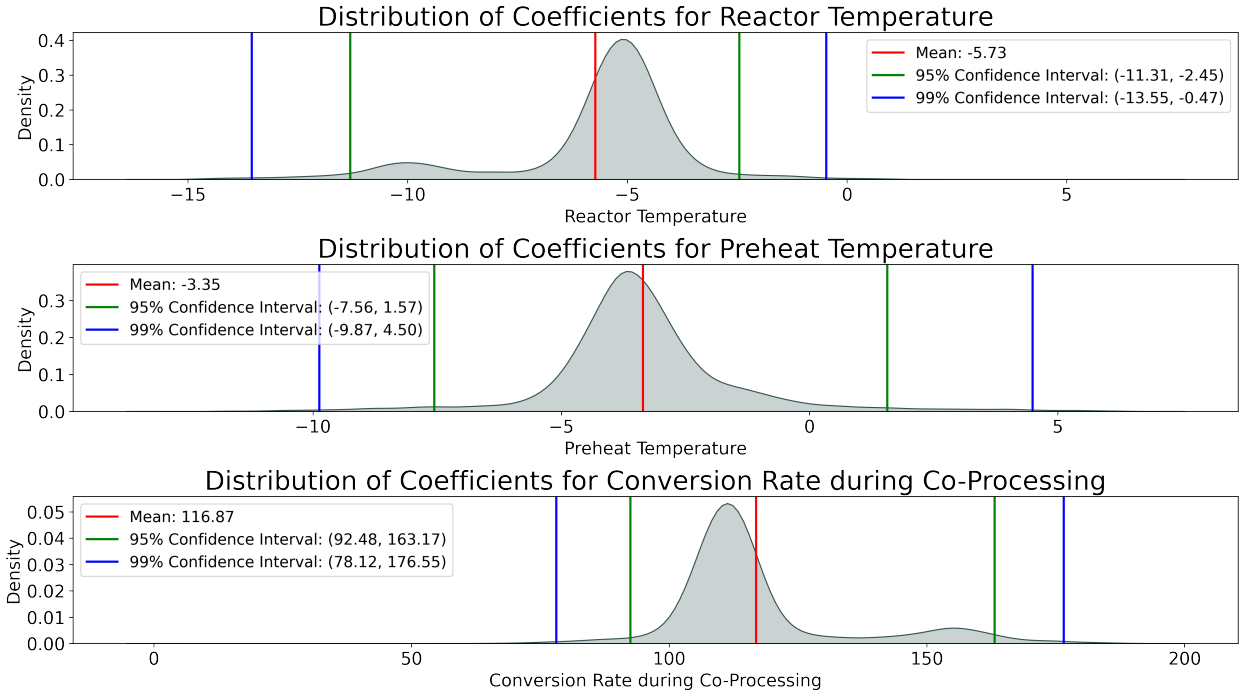

Figure 12: Distribution of other important features coefficients in bootstrapping

Within our study's observed co-processing ratio range of 0 to 21.6%, bio feed generates 59.6% of the LCC compared to fossil feed on a per-unit-volume basis. The mean ratio is 59.3% (with a 95% confidence interval ranging from 56.1% to 61.8%), suggesting that bio feed generates 59.6% of the LCC compared to fossil feed on a per-unit-volume basis. This information provides valuable insights into the relative contributions of fossil feed and bio feed in LCC production during co-processing. We need to emphasize that while the conclusions hold true within our observed range, predictions for co-processing ratios beyond our dataset require further experimental validation.

In studying the Fluid Catalytic Cracking (FCC) process, we found the relationship between reactor temperature and LCC yield to be complex, influenced by factors like feedstock type, catalyst choice, and reactor design. Although one might expect higher temperatures to increase LCC yield, our three-year data analysis mainly shows a negative correlation. However, there are exceptions where higher temperatures benefit LCC yield, highlighting the intricate dynamics involved.

According to our Huber regression model, a rise in Reactor Temperature typically leads to a decrease in LCC by 5.73 units. In terms of renewable LCC, the decrease corresponds to the co-processing ratio multiplied by this number. Interestingly, the 99% confidence interval is between 13.55 and 0.47, indicating a general negative trend, though occasional positive coefficients are observed. Similar complexities are seen with Preheat Temperature and the Conversion Rate during Co-Processing. Figure 12 presents the coefficient distributions for these variables, emphasizing their influence on LCC production.

Figure 13 shows the actual LCC (black line) and predicted LCC (blue line) values on training data and test data, along with the upper and lower bounds of the 95% confidence interval. The predicted renewable LCC, upper and lower bounds of the 95% confidence interval are also provided in the plots.

Since the measurement of renewable carbon with ^{14}C is both expensive and time-consuming, and is limited to only a few laboratories. By using our model, we can predict the ^{14}C (re-

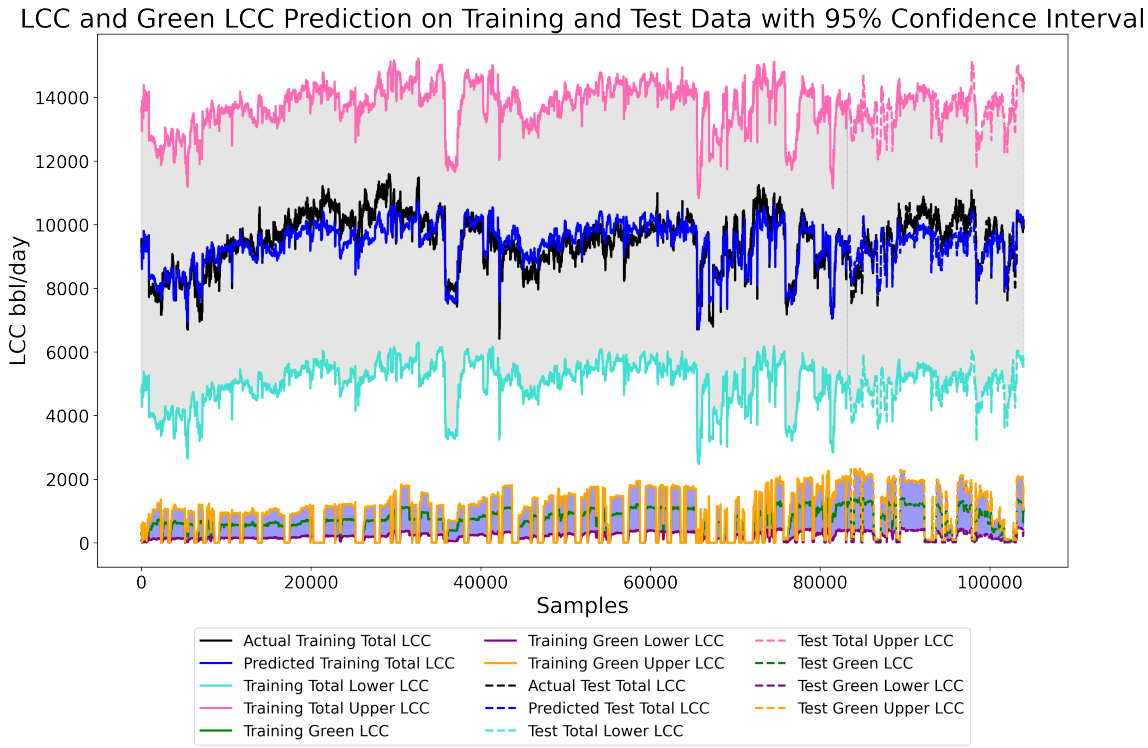

Figure 13: LCC and renewable LCC on training and test data

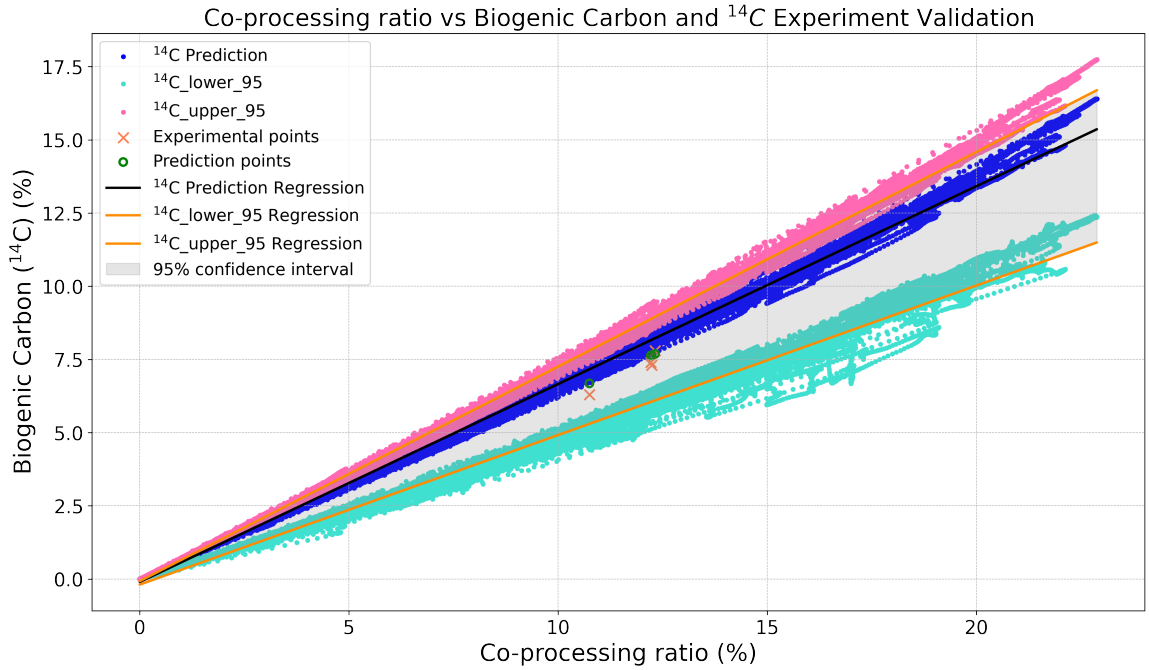

Figure 14: LCC ^{14}C prediction for different co-processing ratio

newable content) in real time without the need for any measurements. The predicted ^{14}C can be defined as follows:

$$Predicted\ ^{14}C = \frac{Renewable\ LCC}{LCC} \tag{9}$$

Figure 14 displays the predicted ^{14}C values and their corresponding 95% confidence interval bounds based on Equation (9). This graph reveals valuable insights into the relationship between co-processing ratio and renewable content. It appears that the renewable content in the product also increased with higher co-processing ratio. This information can help industries optimize their co-processing strategies to achieve desired renewable content levels in their products.

In 2018, Parkland carried out four ^{14}C experiments, each characterized by a unique co-processing ratio. The co-processing ratios were: 10.75%, 12.24%, 12.32%, and 12.22%, with corresponding ^{14}C values of: 6.30%, 7.30%, 7.80%, and 7.40%. In comparison, our AI model, predicting renewable content, yielded ^{14}C values of 6.68%, 7.65%, 7.69%, and 7.64% for the co-processing ratios of 10.75%, 12.24%, 12.32%, and 12.22%, respectively. The experiment results and predictions are shown in Figure 14. The predictions generally align with the experimental results. This consistency highlights our model's accuracy. Furthermore, we calculated the mean ^{14}C values and relative error for different co-processing ratios, and the results are presented in Table 3.

Table 3: Experimental and AI Results of ^{14}C in LCC during Co-processing

Experiment			AI model				
	ratio	^{14}C	ratio(mean)	^{14}C(mean)	^{14}C(lower)	^{14}C(upper)	Error
Sample 1	10.75%	6.30%	10.75%	6.68%	5.24%	8.01%	6.03%
Sample 2	12.24%	7.30%	12.24%	7.65%	6.21%	8.96%	4.79%
Sample 3	12.32%	7.80%	12.32%	7.69%	5.77%	9.40%	1.41%
Sample 4	12.22%	7.40%	12.22%	7.64%	6.09%	8.70%	3.24%

5. Conclusion

In conclusion, this study presents a novel AI-based method for tracking renewable carbon during bio feed co-processing, utilizing large-scale commercial data, interpretable deep neural networks, robust linear regression, and bootstrapping. Our approach efficiently and accurately estimates real-time renewable carbon content in produced liquids, potentially saving oil refineries millions of dollars annually by eliminating the need for expensive ^{14}C measurements. This research holds significant implications for oil refineries transitioning towards low-carbon solutions, as it enables the quantification of renewable carbon for generating credits and optimizing co-processing strategies. Ultimately, this study contributes to the development of sustainable, environmentally friendly energy solutions and encourages further innovation in the renewable energy sector.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgement

We thank Parkland for supporting our work. We thank Mitacs for the financial support.

References

(1) Agency, I. E. Global Energy Review 2021. 2021; https://www.iea.org/reports/global-energy-review-2021.

(2) Agency, I. E. Net Zero by 2050. 2021; https://www.iea.org/reports/net-zero-by-2050.

(3) Su, J.; Cao, L.; Lee, G.; Tyler, J.; Ringsred, A.; Rensing, M.; van Dyk, S.; O'Connor, D.; Pinchuk, R.; Saddler, J. J. Challenges in determining the renewable content of the final fuels after co-processing biogenic feedstocks in the fluid catalytic cracker (FCC) of a commercial oil refinery. *Fuel* **2021**, *294*, 120526.

(4) Fogassy, G.; Thegarid, N.; Schuurman, Y.; Mirodatos, C. From biomass to bio-gasoline by FCC co-processing: effect of feed composition and catalyst structure on product quality. *Energy Environ. Sci.* **2011**, *4*, 5068–5076.

(5) Elliott, D. C. Hydrothermal liquefaction of sludge and biomass residues. **2020**, 117–131.

(6) Elliott, D. C.; Biller, P.; Ross, A. B.; Schmidt, A. J.; Jones, S. B. Hydrothermal liquefaction of biomass: Developments from batch to continuous process. *Bioresource Technology* **2015**, *178*, 147–156.

(7) Biller, P.; Roth, A. Hydrothermal liquefaction: A promising pathway towards renewable jet fuel. **2018**, 607–635.

(8) Badoga, S.; Alvarez-Majmutov, A.; Xing, T.; Gieleciak, R.; Chen, J. Co-processing of Hydrothermal Liquefaction Biocrude with Vacuum Gas Oil through Hydrotreating and Hydrocracking to Produce Low-Carbon Fuels. *Energy & Fuels* **2020**, *34*, 7160–7169.

(9) Su, J.; Cao, L.; Lee, G.; Gopaluni, B.; Siang, L. C.; Cao, Y.; van Dyk, S.; Pinchuk, R.; Saddler, J. Tracking the green coke production when co-processing lipids at a commercial fluid catalytic cracker (FCC): combining isotope 14C and causal discovery analysis. *Sustainable Energy & Fuels* **2022**, *6*, 5600–5607.

(10) Yeh, S.; Witcover, J.; Lade, G. E.; Sperling, D. A review of low carbon fuel policies: Principles, program status and future directions. *Energy Policy* **2016**, *97*, 220–234.

(11) Lammens, T. M. Effect of Various Green Carbon Tracking Methods on Life Cycle

Assessment Results for Fluid Catalytic Cracker Co-processing of Fast Pyrolysis Bio-oil. *Energy & Fuels* **2022**, *36*, 12617–12627.

(12) Jull, A. J. T.; Pearson, C. L.; Taylor, R. E.; Southon, J. R.; Santos, G. M.; Kohl, C. P.; Hajdas, I.; Molnar, M.; Baisan, C.; Lange, T. E.; et al. Radiocarbon Dating and Intercomparison of Some Early Historical Radiocarbon Samples. *Radiocarbon* **2018**, *60*, 535–548.

(13) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer-Verlag: Berlin, Heidelberg, 2006.

(14) LeCun, B. Y. . H., Y. Deep learning. *Nature* **2015**, *521*, 436–444.

(15) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85–117.

(16) Qin, S. J. Process data analytics in the era of big data. *AIChE Journal* **2014**, *60*, 3092–3100.

(17) Yu, F.; Cao, L.; Li, W.; Yang, F.; Shang, C. Feature based causality analysis and its applications in soft sensor modeling. *IFAC-PapersOnLine* **2020**, *53*, 138–143, 21st IFAC World Congress.

(18) Cao, L.; Yu, F.; Yang, F.; Cao, Y.; Gopaluni, R. B. Data-driven dynamic inferential sensors based on causality analysis. *Control Engineering Practice* **2020**, *104*, 104626.

(19) Parkland Parkland announces plans to expand co-processing activities and build British Columbia's largest renewable diesel complex. 2022; `https://www.parkland.ca/en/investors/news-releases/details/parkland-announces-plans-to-expand-co-processing-activities-and-build-british-columbias-largest-renewable-diesel-complex/609`.

(20) Jiang, B.; Luo, Y.; Lu, Q. Maximized Mutual Information Analysis Based on Stochastic Representation for Process Monitoring. *IEEE Transactions on Industrial Informatics* **2019**, *15*, 1579–1587.

(21) Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Communications of the ACM* **2019**, *63*, 68–77.

(22) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **2019**, *116*, 22071–22080.

(23) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

(24) Molnar, C. *Interpretable machine learning*; Lulu. com, 2020.

(25) Stine, R. An Introduction to Bootstrap Methods: Examples and Ideas. *Sociological Methods & Research* **1989**, *18*, 243–291.

(26) Yu, C.; Yao, W. Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation* **2017**, *46*, 6261–6282.