

Article

# Data-Based Sensing and Stochastic Analysis of Biodiesel Production Process

Iftikhar Ahmad <sup>1,\*</sup>, Ahsan Ayub <sup>2</sup>, Uzair Ibrahim <sup>1</sup>, Mansoor Khan Khattak <sup>3</sup>  
and Manabu Kano <sup>4</sup>

<sup>1</sup> Department of Chemical and Materials Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan; uibrahim\_che07@scme.nust.edu.pk

<sup>2</sup> US Pakistan Center for Advanced Studies in Energy, National University of Sciences and Technology, Islamabad 44000, Pakistan; ahsan\_che06@scme.nust.edu.pk

<sup>3</sup> Department of Agricultural Mechanization, The University of Agriculture Peshawar, Peshawar 25000, Pakistan; mansoorkhankhattak@yahoo.com

<sup>4</sup> Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan; manabu@human.sys.i.kyoto-u.ac.jp

\* Correspondence: iftikhar.salarzai@scme.nust.edu.pk; Tel.: +92-5190855108

Received: 6 October 2018; Accepted: 26 November 2018; Published: 25 December 2018

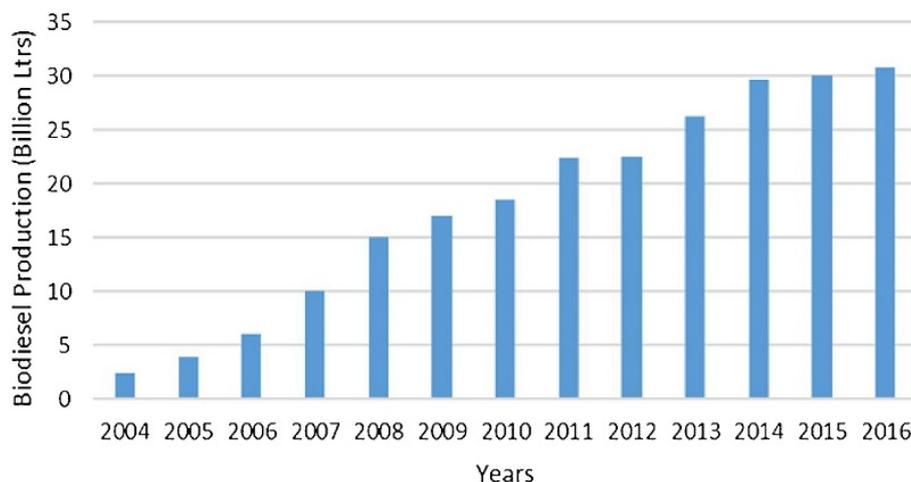


**Abstract:** Biodiesel production is a field of outstanding prospects due to the renewable nature of its feedstock and little to no overall CO<sub>2</sub> emissions to the environment. Data-based soft sensors are used in realizing stable and efficient operation of biodiesel production. However, the conventional data-based soft sensors cannot grasp the effect of process uncertainty on the process outcomes. In this study, a framework of data-based soft sensors was developed using ensemble learning method, i.e., boosting, for prediction of composition, quantity, and quality of product, i.e., fatty acid methyl esters (FAME), in biodiesel production process from vegetable oil. The ensemble learning method was integrated with the polynomial chaos expansion (PCE) method to quantify the effect of uncertainties in process variables on the target outcomes. The proposed modeling framework is highly accurate in prediction of the target outcomes and quantification of the effect of process uncertainty.

**Keywords:** biodiesel; machine learning; ensemble learning; boosting; uncertainty analysis; polynomial chaos expansion

## 1. Introduction

Extensive use of fossil fuels is causing environmental issues, i.e., global warming and pollution, and depletion of energy resources [1]. These challenges drive the quest for exploring alternative energy resources that can help in the reduction of fossil fuels consumption and environmental impact. Bioenergy is one of the viable alternatives that is shifting the paradigm from conventional fuels to more sustainable resources. Biodiesel is one of the major bio-based fuels, and a drastic increase in its production has been reported in the last two decades, as shown in Figure 1 [2,3]. Efficient design and operation of biodiesel production process are investigated to minimize the consumption of raw materials and utilities and also produce a high quality product. Data-based soft sensors are used to realize stable operation of the biodiesel production. The data-based soft sensors are more efficient than the model-based soft sensors in capturing the non-linearity of complex processes and prediction of desired outcomes; an extensive review on data-based soft sensors can be found in [4]. Their applications in the biodiesel production process include online prediction, optimization, and control.



**Figure 1.** Worldwide production of biodiesel [3].

Studies on online prediction mostly take kinematic viscosity, density, and cetane number of biodiesel as their target outcomes. For instance, Meng et al. (2013) developed an artificial neural networks (ANN) model to predict the biodiesel kinematic viscosity [5]. Rocabrundo-Valdés et al. (2015) used an ANN model to predict density, dynamic viscosity, and cetane number of biodiesel [6]. Mostafaei et al. (2016) evaluated the efficiency of the response surface methodology (RSM) and neuro-fuzzy inference system (ANFIS) in modeling the yield achieved in an ultrasonic reactor [7]. Miraboutalebi et al. (2016) used random forest and ANN to predict cetane number of biodiesel [8]. Raman et al. (2018) developed an ANN model for estimation of densities of vegetable oil based ethyl esters biodiesel [9].

The concept of data-based soft sensing is also extended to parametric analysis, optimization, and control of the biodiesel production process. Wali et al. (2013) investigated online intelligent controllers for real-time temperature control in an advanced biodiesel microwave reacting system [10]. Fayyazi et al. (2014) determined optimum temperature, catalytic concentration, and reaction time in biodiesel production processes using genetic algorithm [11]. Sikorski et al. (2016) applied polynomial and high-dimensional model representation (HDMR) fitting to analyze the effect of process variables on the heat duties of equipment in a biodiesel production process [12]. Cheng et al. (2016) proposed a genetic algorithm-based evolutionary support vector machine (GA-ESVM) to get optimum mixture properties for higher biodiesel yield [13].

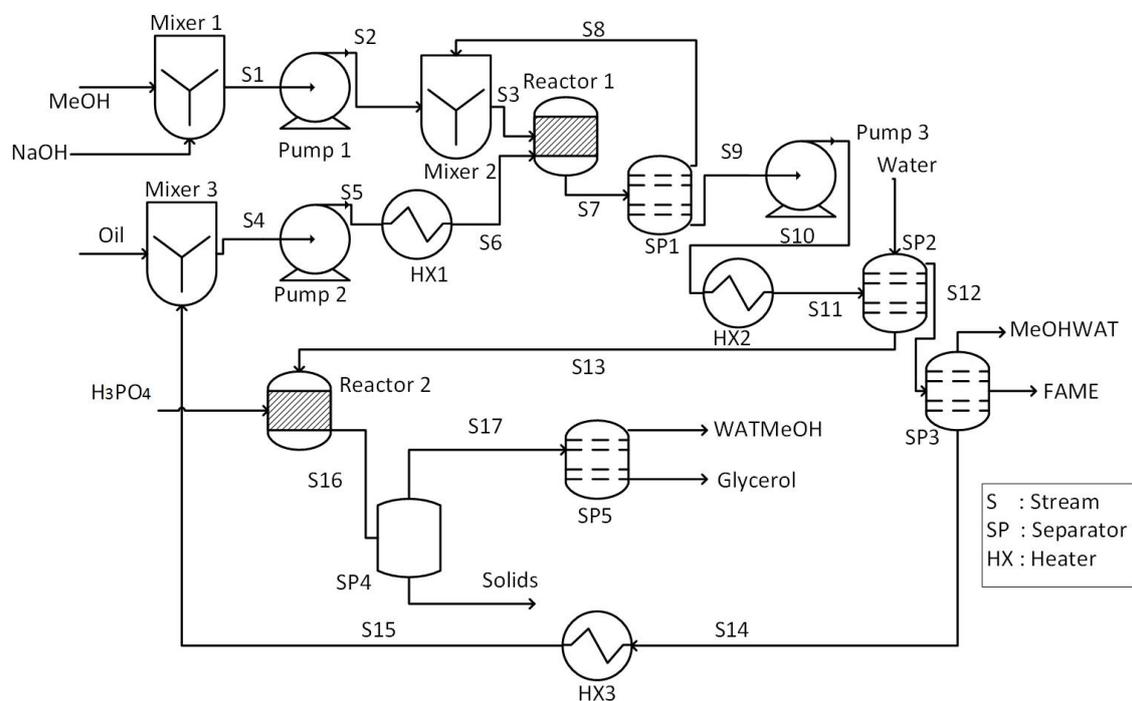
The efficiency of the data-based soft sensing methods is deteriorated due to uncertainty in process conditions. To realize a more robust sensing system, some uncertainty quantification mechanism should be incorporated in the soft sensor framework [14]. Uncertainty analysis is used to quantify the impact of uncertainty in model input on the model output [15]. It has been the focus of research and several methods are reported in the literature [16]. Scenario analysis, multiple model simulation methods, inverse modeling method, sensitivity analysis, and sampling-based method are commonly used for uncertainty analysis. The scenario analysis quantifies the impact of uncertainty associated with future developments on the model performance or relevance [17]. The multiple model simulation evaluates several modeling structures to determine the overall impact of structural uncertainty on the model performance [18,19]. The inverse modeling method assesses the effect of uncertainty in process parameters on the model performance and helps in optimizing the parameters [20,21]. Sensitivity analysis (SA) derives a hierarchy of model input in terms of their contribution to the model output uncertainty [22]. The sample-based uncertainty analysis methods, i.e., Monte Carlo and polynomial chaos expansion (PCE), quantify the collective impact of uncertainty in model input on its output [15].

In this study, an integrated framework of data-based soft sensing and uncertainty analysis was proposed. Data-based soft sensors were developed using ensemble learning method, i.e., boosting, to predict composition, quantity, and quality of fatty acid methyl esters (FAME) in the outlet streams of biodiesel production process; cetane number of the FAME was used as a quality parameter. The cetane number relates to the ignition delay time of a fuel and is applied to alternative diesel fuels such as biodiesel and its components [23]. An increase in cetane number reduces the ignition delay period and stabilizes running of the engine. However, an excessive rise in cetane number causes too much reduction in the ignition delay. As a result, the fuel does not have proper time to spread into the combustion chamber and performance of the engine decreases. Prediction of composition helps in maintaining desired level of mole fraction of components while prediction of cetane number assists in realizing high quality of FAME. Prediction of quantity, i.e., flow rate, of FAME is important for evaluating conversion efficiency of the process. Polynomial chaos expansion (PCE) method was incorporated into the development of the soft sensor to quantify the effect of process uncertainties on the target outcomes, i.e., composition, quantity, and quality of FAME. PCE is computationally less expensive than other sampling method such as Monte Carlo. The incorporation of PCE transformed the prediction of the ensemble model from deterministic to stochastic format where the effect of process uncertainties is visualized in the form of predictive distributions.

Section 2 explains the process description for biodiesel production followed by modeling methods described in Section 3. Section 4 outlines the proposed method. Section 5 shows the results and discussion, while Section 6 concludes the work.

## 2. Process and Data Description

A process flow diagram of biodiesel production from vegetable oil is shown in Figure 2.



**Figure 2.** Process flowsheet of biodiesel production from vegetable oil.

It is a base catalyzed process that utilizes NaOH as a catalyst. The main steps in the process involve transesterification reaction, biodiesel separation, glycerol separation, and recovery of methanol. Vegetable oil feed along with recycled oil is heated to reaction temperature using heater (HX1) and then fed to the reactor (Reactor 1). In addition, a mixture of NaOH and methanol (MeOH) is also

charged into the reactor where transesterification reaction occurs. Effluent of the reactor is separated into two component by the first separator (SP1); excess methanol is recovered as the top stream and recycled, while bottom stream, i.e., fatty acid methyl esters (FAME), is sent to the second separator (SP2) for further purification.

Water washing is performed in SP2 to remove glycerol, catalyst, and unconverted methanol from FAME. The top stream (S12) is sent to the third separator (SP3) where unconverted oil is separated to get further purified FAME as a product. The bottom stream of SP2 is fed to the other reactor (Reactor 2) where phosphoric acid is used to neutralize the stream. Solids are removed from the effluent of Reactor 2 by filtration in SP4. Then, another separator (SP5) is used to remove water from the top stream of SP4 and produce pure glycerol as a by-product.

### 3. Fundamentals of Modeling and Analysis Methods

#### 3.1. Soft-Sensor Development

Boosting, which is an ensemble learning technique, is adopted in this study for the soft sensor development. Boosting is based on the idea of developing a robust model by combining several weak models [24]. The concept of boosting is demonstrated in Figure 3 [25]. The models are developed in a series of rounds where the focus on incorrectly predicted target samples is increased with the help of increasing their respective weight. Several boosting mechanism are developed on the basis of variation in their methods. Least Squares Boosting (LSBoost) was adopted in this study [26].

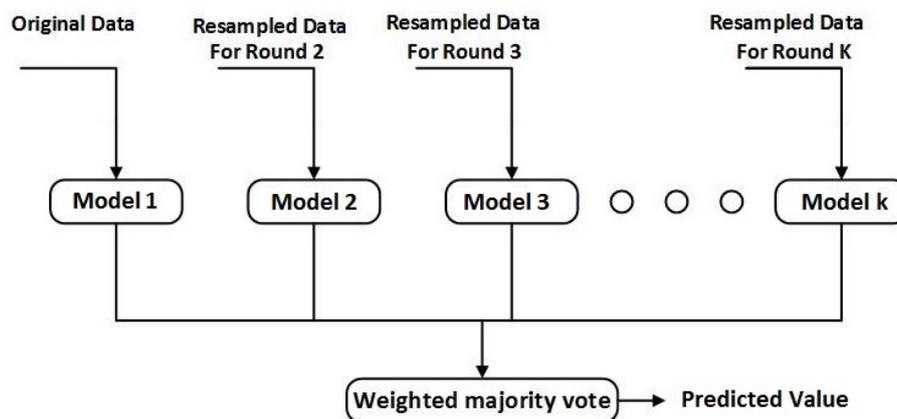


Figure 3. The framework of boosting.

#### 3.2. Uncertainty Analysis

Uncertainty analysis quantifies the accumulated impact of uncertainties of all input variables of a model [27]. Uncertainty has several dimensions, i.e., location, level, and nature; an extensive review of dimensions of uncertainty in process modeling was done by Ahmad et al. [14]. There are several methods available for uncertainty analysis of a process model. This study utilized sampling-based methods, i.e., polynomial Chaos Expansion (PCE), for uncertainty analysis of the process. The sampling-based uncertainty analysis helps in determining collective impact of uncertainties in all process variables on the process output.

In PCE, a random variable  $x$  is represented as a function ( $f()$ ) of another random variable ( $\xi$ ) [28,29]:

$$x = f(\xi) \quad (1)$$

The PCE seeks an appropriate function  $f()$ , by describing  $x$  through deterministic and stochastic components:

$$\mathbf{x} = f(\xi) = \sum_{i=1}^{\infty} \alpha_i \psi_i(\xi) \quad (2)$$

where  $\alpha_i$  and  $\psi_i$  are the deterministic and the stochastic components, respectively.  $\psi_i$  is a polynomial that satisfy the following condition:

$$\langle \psi_j, \psi_k \rangle = \int \psi_j(\xi) \psi_k(\xi) p_{\xi}(\xi) d\xi = 0, j \neq k \quad (3)$$

where  $\langle \psi_j, \psi_k \rangle$  is the inner product of  $\psi_j$  and  $\psi_k$ , and  $p_{\xi}$  is the probability distribution function (PDF) of  $\xi$ .

To implement the PCEs, the mode strengths should be estimated by intrusive or non-intrusive (black box) methods [30,31]. In the current study, we implemented a non-intrusive approach where an ensemble model was used as a black-box system.

#### 4. Proposed Modeling and Analysis Framework

The proposed modeling strategy is shown in Figure 4 and summarized as follows:

1. *Data generation*: Data are generated by inserting variations in the steady state values of the process model through interfacing of MATLAB<sup>®</sup>, Excel<sup>®</sup> and Aspen<sup>®</sup> environments. Lists of process input and output variables used for the model development are shown in Tables 1 and 2, respectively.
2. *Soft-sensor Design*: The generated data are used to develop the soft-sensors through the ensemble learning method. The number of decision trees, i.e. weak learners, in the ensemble models is optimized.
3. *PCE based uncertainty analysis*: The ensemble model developed in Step 2 is used within the PCE framework. PCE level and the number of terms are optimized. A uniform uncertainty in all input variables is assumed and PCE based random variables are generated for each of the input variables. The PCE based generated random variables are fed to the ensemble model and predictive distributions of respective outputs are obtained.

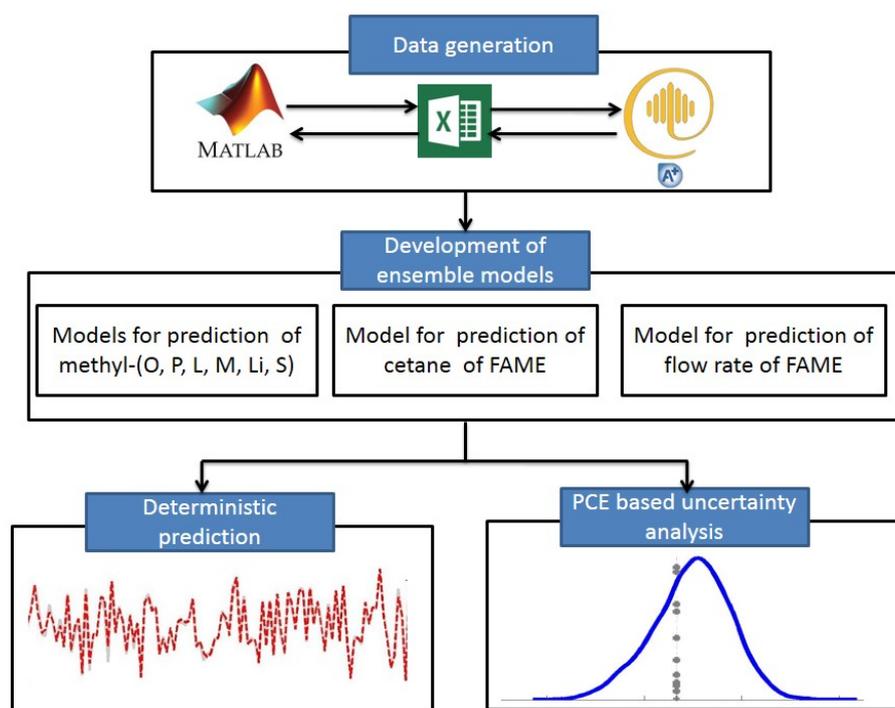


Figure 4. Proposed modeling framework.

**Table 1.** Process input variables.

No.	Names	No.	Names
1	MeOH steam's flow rate	26	Water steam's temperature
2	MeOH steam's temperature	27	Water steam's pressure
3	MeOH steam's pressure	28	Water stream's flow rate
4	NaOH steam's pressure	29	H <sub>3</sub> PO <sub>4</sub> steam's temperatre
5	NaOH steam's temperature	30	H <sub>3</sub> PO <sub>4</sub> steam's pressure
6	NaOH steam's total flow rate	31	H <sub>3</sub> PO <sub>4</sub> steam's total flow rate
7	NaOH percentage in total fowrate	32	H <sub>3</sub> PO <sub>4</sub> percentage in total flow rate
8	Water percentage in total flow rate	33	Water percentage in total flow rate
9	Oil stream's temperature	34	Pump 1 pressure
10	Oil stream's pressure	35	Pump 2 pressure
11	Oil stream's total flow rate	36	HX1 temperature
12	Triolein flow rate	37	Reactor1 temperature
13	Trimyristin flow rate	38	Reactor1 pressure
14	Triplamitin flow rate	39	Reactor1 residence time
15	Polyphenyl sulfide flow rate	40	SP1 basis
16	Diphenyloxazole flow rate	41	SP1 reflux rate
17	Oleo-palmitostearin flow rate	42	Pump 3 pressure
18	Matrix metalloproteinase flow rate	43	HX2 temperature
19	Dipalmitoyl linoleoyl glycerol flow rate	44	SP3 basis
20	Palmitoyl dioleoylglycerol flow rate	45	SP3 reflux frate
21	Linoleoyl oleoyl palmitoyl glycerol flow rate	46	Reactor2 temperature
22	Dioleoyl stearo glycerol flow rate	47	Reactor2 pressure
23	Dilinoleoyl stearoyl glycerol flow rate	48	SP5 basis
24	Monolauroyl dioleoyl glycerol flow rate	49	SP5 reflux rate
25	Dipalmitoyl glycerol flow rate		

**Table 2.** Process output variables used for ensemble learning.

No.	Names	Values	Units
1	Oleic Acid Methyl Ester(Methyl O)	397.8	kg/h
2	Palmitic Acid Methyl Ester(Methyl P)	518.9	kg/h
3	Myristic Acid Methyl Ester(Methyl M)	22.57	kg/h
4	Linoleic Acid Methyl Ester(Methyl Li)	79.68	kg/h
5	Stearic Acid Methyl Ester(Methyl S)	29.67	kg/h
6	FAME	1050	kg/h
7	Cetane Number	64.78	

MATLAB<sup>®</sup>, Excel<sup>®</sup>, and Aspen<sup>®</sup> were interfaced to generate 525 data samples. In the ensemble modeling, the weak learners and their optimized number for the respective models are shown in Table 3. In the PCE based method, Hermite function was used for a level of 6 and initial 20 terms.

**Table 3.** Summary of soft sensors predictions.

No.	Output Variable	Weak Learners	% Accuracy	RMSE Value	SSE Value
1	Methyl Li	1200	98.779	$3.954 \times 10^{-6}$	$1.61 \times 10^{-9}$
2	Methyl O	1200	99.153	$2.391 \times 10^{-5}$	$5.89 \times 10^{-8}$
3	Methyl M	400	98.901	$6.906 \times 10^{-6}$	$4.912 \times 10^{-6}$
4	Methyl P	600	99.394	$1.488 \times 10^{-5}$	$2.279 \times 10^{-8}$
5	Methyl S	2000	98.337	$4.188 \times 10^{-6}$	$1.807 \times 10^{-9}$
6	FAME flow rate	800	99.527	0.6025	37.99
7	Cetane Number	650	99.531	0.0396	0.1407

## 5. Results and Discussion

This section covers the application of the proposed integrated framework of data-based soft sensing and uncertainty analysis to the biodiesel production process from vegetable oil.

Seven ensemble models (soft sensors) were developed, one for flow rate of FAME, one for cetane number, and one each for prediction of the mole fractions of the components, i.e., Methyl-Li, Methyl-M, Methyl-O, Methyl-S and Methyl-P.

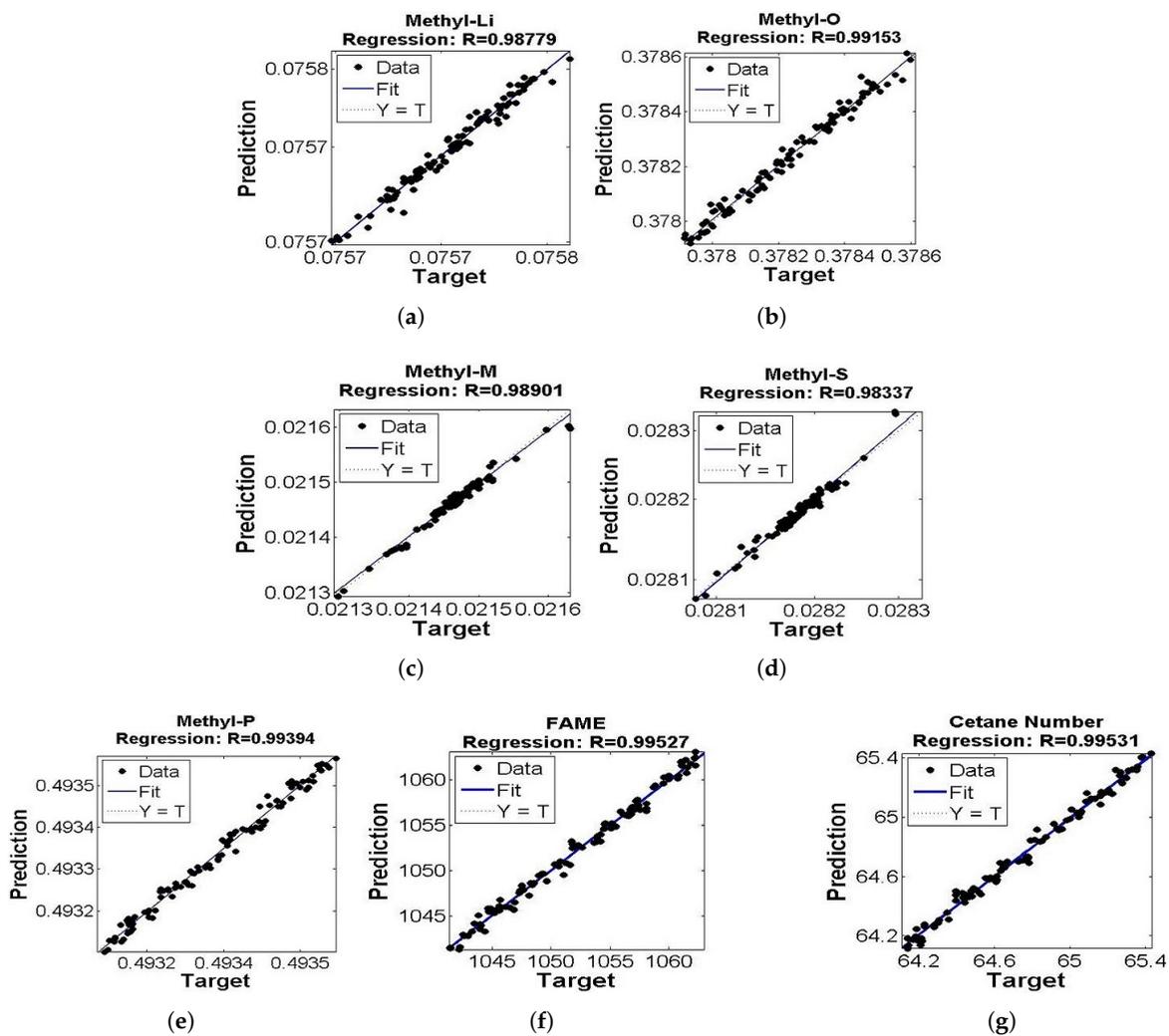
Cetane number for biodiesel was calculated using the following equation:

$$CN = \sum_t X_t \times CN_t \quad (4)$$

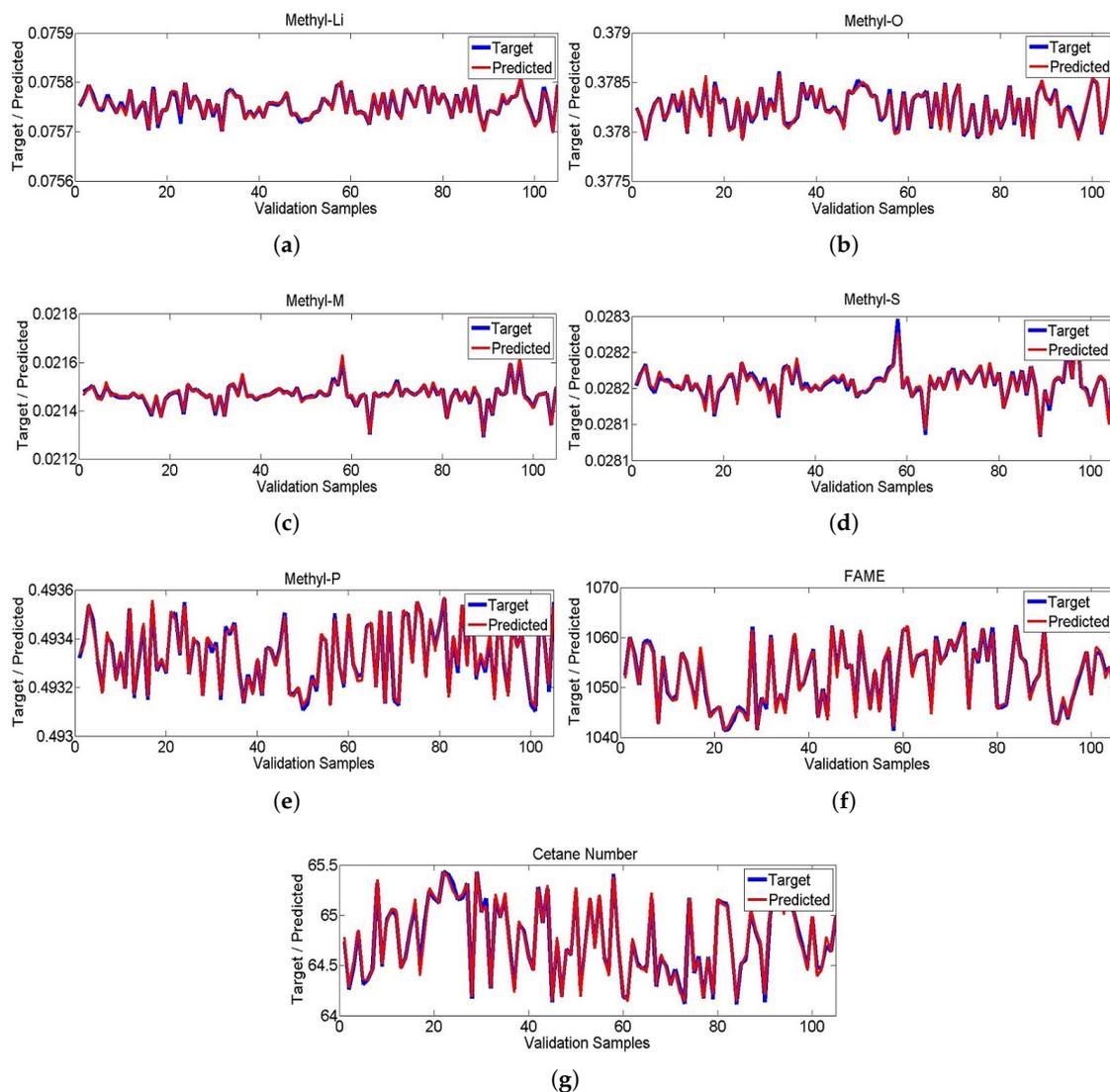
where  $CN$  represents cetane number of biodiesel;  $t$  refer to the type of methyl ester component, i.e., Methyl-Li, Methyl-M, Methyl-O, Methyl-S and Methyl-P; and  $X$  represents mass fraction.

A total of 525 datasets comprising the input and output data were generated for each model; data were divided into training (80%) and validation (20%) sets. Training sets were used for model development, while validation sets were used to evaluate models' accuracy.

Performance of the soft sensors is plotted in Figures 5 and 6.



**Figure 5.** (a–g) Regression performance of ensemble model of Methyl-Li, Methyl-O, Methyl-M, Methyl-S, Methyl-P, Fame and Cetane number.



**Figure 6.** (a–g) Target and predicted values of Methyl-Li, Methyl-O, Methyl-M, Methyl-S, Methyl-P, FAME and Cetane number.

A correlation coefficient between the target and predicted values of soft sensors was used to evaluate prediction accuracy. The correlation coefficients of soft sensor for Methyl-Li, Methyl-O, Methyl-M, Methyl-P, and Methyl-S components were 0.9877, 0.9915, 0.9890, 0.9939 and 0.9833, respectively. Root-mean-square errors (RMSE) for Methyl-Li, Methyl-O, Methyl-M, Methyl-P and Methyl-S components were  $3.954 \times 10^{-6}$ ,  $2.391 \times 10^{-5}$ ,  $6.906 \times 10^{-6}$ ,  $1.488 \times 10^{-5}$ , and  $4.188 \times 10^{-6}$ , respectively. Squared errors of prediction (SSE) for Methyl-Li, Methyl-O, Methyl-M, Methyl-P and Methyl-S components were  $1.61 \times 10^{-9}$ ,  $5.89 \times 10^{-8}$ ,  $4.912 \times 10^{-6}$ ,  $2.279 \times 10^{-8}$ , and  $1.807 \times 10^{-9}$ , respectively.

The correlation coefficient of flow rate of FAME was 0.9952, as shown in Figures 5 and 6. RMSE and SSE values for FAME were 0.6025 and 37.99, respectively. Similarly, the correlation coefficient for cetane number was 0.9953 as shown in Figures 5 and 6. RMSE and SSE were 0.0396 and 0.1407, respectively.

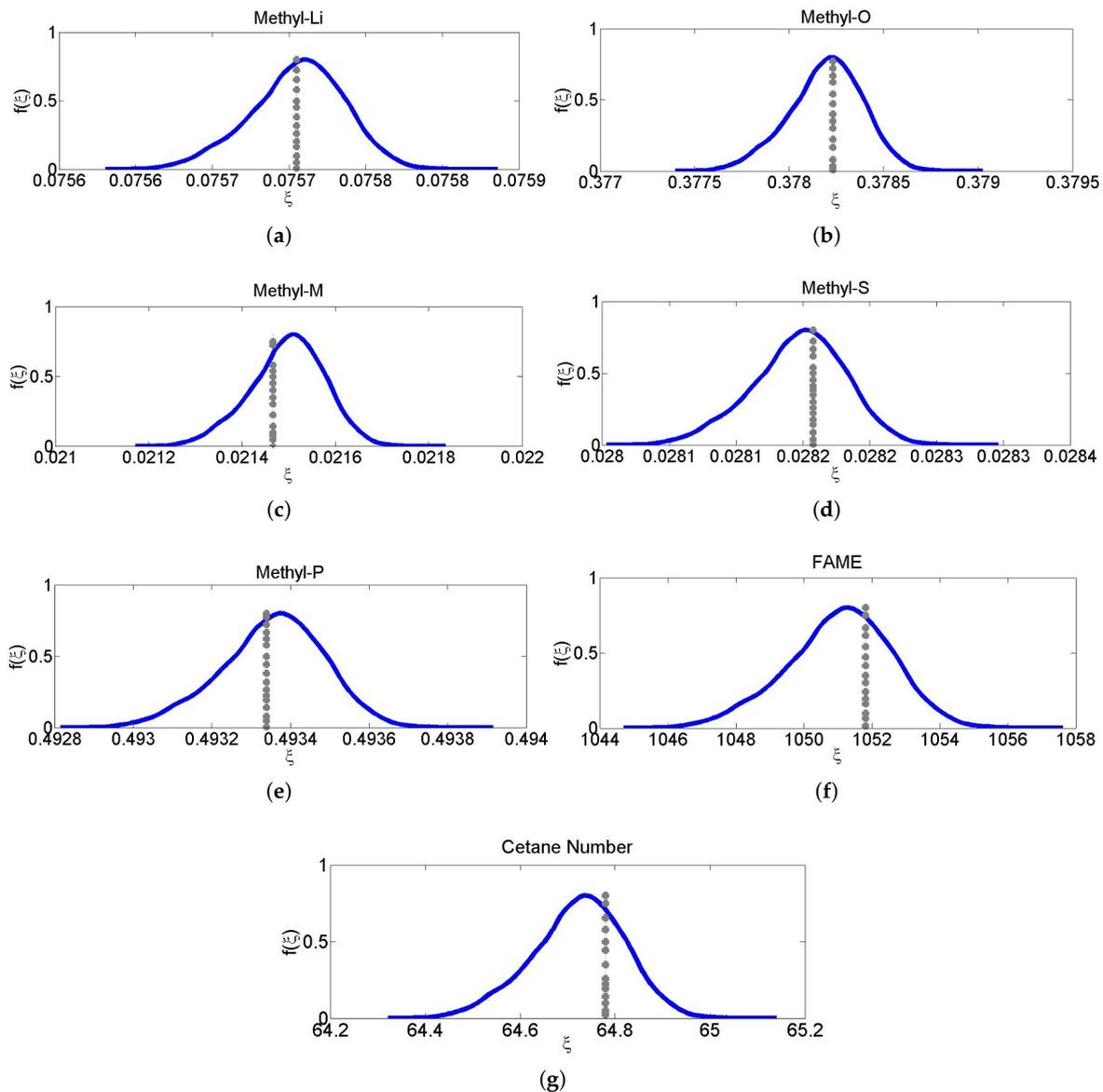
Table 3 shows a summary of prediction accuracy (correlation coefficient), RMSE and SSE values for all the seven soft sensors. Actual values of target variables in validation datasets and percent errors in their corresponding predicted values are shown in Table 4.

**Table 4.** Actual values of target outcomes of soft sensors and errors exhibited during their prediction.

Actual Values							Error (%)						
P	Li	O	M	S	CN	FAME	P	Li	O	M	S	CN	FAME
0.49334	0.07575	0.37823	0.02147	0.02821	64.77983	1051.82387	0.00352	0.00336	0.00323	0.05590	0.01239	0.07194	0.05856
0.49337	0.07577	0.37815	0.02149	0.02822	64.28166	1059.97524	0.00346	0.00203	0.00479	0.00775	0.00722	0.03378	0.00347
0.49354	0.07579	0.37793	0.02150	0.02823	64.52924	1055.90848	0.00104	0.00039	0.00378	0.01603	0.01017	0.07744	0.05311
0.49344	0.07576	0.37814	0.02145	0.02821	64.85225	1050.64924	0.00793	0.01181	0.00439	0.04947	0.01045	0.01915	0.01297
0.49331	0.07575	0.37829	0.02145	0.02820	64.36030	1058.68015	0.00235	0.00113	0.00015	0.02177	0.01737	0.07569	0.04245
0.49318	0.07575	0.37833	0.02152	0.02822	64.35063	1058.83926	0.00194	0.00751	0.00459	0.08273	0.03344	0.00743	0.05034
0.49343	0.07577	0.37814	0.02146	0.02821	64.46647	1056.93662	0.00450	0.01071	0.00623	0.05596	0.00571	0.00400	0.00197
0.49337	0.07576	0.37820	0.02147	0.02821	65.35289	1042.60063	0.00331	0.00485	0.00585	0.03767	0.01174	0.04987	0.02124
0.49323	0.07574	0.37838	0.02146	0.02820	64.51843	1056.08529	0.00121	0.00075	0.00278	0.00945	0.00470	0.03943	0.01381
0.49335	0.07575	0.37824	0.02146	0.02820	64.91510	1049.63203	0.00247	0.00833	0.00870	0.00704	0.00951	0.05944	0.07687
0.49319	0.07573	0.37842	0.02146	0.02820	65.06075	1047.28217	0.00027	0.00619	0.01187	0.00108	0.00013	0.02891	0.02542
0.49352	0.07578	0.37800	0.02148	0.02822	65.04558	1047.52648	0.00036	0.00243	0.00264	0.01259	0.00175	0.07035	0.08327
0.49317	0.07573	0.37844	0.02146	0.02820	64.46980	1056.88189	0.00317	0.00209	0.00205	0.01450	0.00381	0.01719	0.00665
0.49332	0.07575	0.37825	0.02147	0.02821	64.57717	1055.12474	0.00224	0.00774	0.00933	0.02765	0.00063	0.01173	0.00227
0.49346	0.07576	0.37815	0.02143	0.02820	64.87258	1050.32008	0.00256	0.00124	0.00737	0.03615	0.01075	0.00821	0.03637
0.49316	0.07571	0.37857	0.02139	0.02817	65.16236	1045.64917	0.00273	0.00275	0.01595	0.04569	0.03807	0.11464	0.10758
0.49356	0.07579	0.37797	0.02147	0.02822	64.39687	1058.07882	0.00281	0.00591	0.00420	0.04087	0.00622	0.16337	0.14401
0.49327	0.07572	0.37847	0.02138	0.02817	64.82592	1051.07598	0.00120	0.00937	0.00967	0.00119	0.01301	0.14071	0.14541
0.49332	0.07574	0.37829	0.02145	0.02820	65.26961	1043.93092	0.00181	0.00229	0.00010	0.00044	0.00617	0.04575	0.01654
0.49327	0.07574	0.37832	0.02147	0.02820	65.19200	1045.1737	0.00167	0.00131	0.00610	0.00296	0.00477	0.03743	0.01871
0.49351	0.07578	0.37801	0.02148	0.02822	65.12934	1046.17934	0.00443	0.00215	0.00833	0.01145	0.00528	0.00977	0.01049
0.49348	0.07578	0.37803	0.02149	0.02822	65.42734	1041.41424	0.00570	0.01090	0.00177	0.05628	0.01134	0.00765	0.00286
0.49334	0.07573	0.37835	0.02140	0.02818	65.37571	1042.23679	0.00163	0.02018	0.00616	0.04574	0.02140	0.04538	0.05826
0.49353	0.07580	0.37792	0.02151	0.02824	65.24007	1044.40363	0.00295	0.00026	0.00585	0.02183	0.00675	0.12215	0.10655
0.49325	0.07575	0.37831	0.02148	0.02821	65.16601	1045.59063	0.00001	0.00042	0.00527	0.00645	0.01937	0.01289	0.04896

Note: P, Li, O, M and S refer to Methyl-P, Methyl-Li, Methyl-M, Methyl-S and Methyl-P, respectively.

The non-intrusive PCE based predictive distributions of Methyl-Li, Methyl-O, Methyl-M, Methyl-P, and Methyl-S components, and FAME flow rate and cetane number are shown in Figure 7; the dotted gray lines show actual values of the respective components, while the blue lines show the PCE based predictive distributions. Actual values refer to the steady state values of the target variables of the process model. Table 5 shows mean absolute deviation percent (MADP) in the target variables of some selected validation datasets for 1% uncertainties in actual values of all input variables.



**Figure 7.** (a–g) PCE based Predicted distribution of measured composition, cetane number and flow rate of FAME (blue lines), measured values of composition, cetane number and flow rate of fame (grey dotted line).

**Table 5.** Mean absolute deviation percent (MADP) of output variables, i.e., compositions, flow rate and cetane number, from their respective measured values.

P	Li	O	M	S	CN	FAME
0.16552	0.27808	0.34019	2.40974	0.82984	1.08191	1.05554
0.16577	0.27457	0.31979	2.38871	0.83584	1.05563	0.95637
0.16292	0.27772	0.31339	2.39248	0.83512	1.02864	0.91446
0.16308	0.27550	0.31750	2.34861	0.84314	1.04743	1.03876
0.16093	0.27067	0.33528	2.41099	0.82704	1.08012	1.00535
0.16059	0.27486	0.31578	2.45583	0.83230	0.99584	0.96712
0.16267	0.27954	0.31831	2.37411	0.83245	1.03889	0.95274
0.16841	0.27268	0.31927	2.40162	0.83223	0.96045	0.96516
0.16376	0.27824	0.32129	2.37550	0.82250	0.99785	0.93083
0.16149	0.27109	0.33075	2.42211	0.83674	0.93986	0.99403
0.16636	0.26695	0.32182	2.41314	0.82858	1.04694	1.00870
0.16065	0.26844	0.32060	2.40208	0.82506	0.95842	1.01706
0.16265	0.27800	0.32840	2.46668	0.82138	1.01479	1.01629
0.16915	0.28244	0.32207	2.37457	0.83700	0.99861	0.98543
0.16623	0.27555	0.31957	2.33797	0.83614	1.07821	1.06566
0.16651	0.27631	0.32403	2.36810	0.84555	1.06042	1.02553
0.15890	0.27896	0.32217	2.39600	0.81433	0.96692	1.10511
0.16413	0.26931	0.31606	2.38986	0.82274	1.03271	0.97631
0.16124	0.27469	0.33022	2.41911	0.84660	1.03330	1.00750
0.16817	0.28308	0.32721	2.42816	0.83586	1.04583	1.00173
0.16146	0.27422	0.31547	2.40128	0.83253	0.98816	1.00466
0.16226	0.27761	0.32194	2.43287	0.83526	0.91761	0.86320
0.16428	0.27386	0.31823	2.40904	0.84382	0.98527	0.92505
0.15903	0.27715	0.31880	2.36198	0.82594	1.05677	0.98890
0.16761	0.27635	0.33213	2.41000	0.83058	1.09828	1.08529

Note: P, Li, O, M and S refer to Methyl-P, Methyl-Li, Methyl-M, Methyl-S, and Methyl-P, respectively.

For Methyl-Li, Methyl-O, Methyl-M, Methyl-P, and Methyl-S components, MADP values were 0.27479, 0.32227, 2.41208, 0.1651, and 0.82135, respectively. For FAME flow rate and cetane number, MADP values were 0.96546 and 0.97013, respectively.

The boosting framework adopted in this study demonstrated high efficiency in predicting the desired outcomes. It is worth mentioning that boosting based soft sensors outperform other soft sensors based on single data-driven model such as ANN and other ensemble learning models such as RF [32,33]. The current soft sensing framework is more intuitive because the outputs cover all features of the product, i.e., quantity, quality and components affecting the quality. The multi-layer estimation of the process outputs promotes the efficiency of the process operation. In development of the soft sensors, many input variables were used, which enables them (soft sensors) to capture the actual dynamics of the process better than the soft sensors based on lesser number of input variables. Although the data used in the development of the soft sensors were extracted from an Aspen PLUS<sup>®</sup> model, the framework can be replicated for the process of a real biodiesel production plant where the raw material is vegetable oil.

The assumption of 1% uncertainty in all input variables was not based on reference information from plant operation but it helps in establishing a quantitative relation between the input uncertainty and their collective impact on the process outputs. MADP was used to quantify the deviation in the process output from their actual values. The deviation determined through the proposed framework can help in developing a control system for ensuring high yield and quality in a biodiesel production plant. In that context, parametric analysis of the process would be needed to identify the variables to be manipulated for maintaining desired values of the process outputs.

## 6. Conclusions

In this study, a data-based soft sensing mechanism was developed to predict composition, flow rate, and cetane number of fatty acid methyl esters (FAME). The non-intrusive polynomial chaos expansion (PCE) method was integrated in the soft sensors framework to quantify the effect of uncertainty on the soft sensors outcomes. A separate model (soft sensor) was developed for each of the components, flow rate and cetane. Prediction accuracies of Methyl-Li, Methyl-M, Methyl-O, Methyl-S, Methyl-P, FAME flow rate, and cetane number were 0.9877, 0.9890, 0.9915, 0.9833, 0.9939, 0.9952 and 0.9953, respectively. For 1% uncertainty in all input variables of the soft sensors, mean absolute deviation percent (MADP) values of 0.27479, 0.32227, 2.41208, 0.1651, 0.82135, 0.96546, and 0.97013 were noticed in the predicted values of Methyl-Li, Methyl-O, Methyl-M, Methyl-P, Methyl-S, FAME flow rate, and cetane number, respectively. The sensors are highly accurate in prediction and uncertainty quantification which make them suitable for real time applications.

**Author Contributions:** Conceptualization, I.A.; Software, I.A. and A.A.; Formal analysis, I.A. and A.A.; Methodology, I.A.; Supervision, I.A.; Writing—review & editing, A.A., U.I., M.K.K. and M.K.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Atadashi, I.; Aroua, M.; Aziz, A.A.; Sulaiman, N. Refining technologies for the purification of crude biodiesel. *Appl. Energy* **2011**, *88*, 4239–4251. [[CrossRef](#)]
2. Atadashi, I.; Aroua, M.; Aziz, A.A.; Sulaiman, N. Production of biodiesel using high free fatty acid feedstocks. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3275–3285. [[CrossRef](#)]
3. Jain, M.; Chandrakant, U.; Orsat, V.; Raghavan, V. A review on assessment of biodiesel production methodologies from Calophyllum inophyllum seed oil. *Ind. Crop. Prod.* **2018**, *114*, 28–44. [[CrossRef](#)]
4. Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33*, 795–814. [[CrossRef](#)]
5. Meng, X.; Jia, M.; Wang, T. Neural network prediction of biodiesel kinematic viscosity at 313 K. *Fuel* **2014**, *121*, 133–140. [[CrossRef](#)]
6. Rocabruno-Valdés, C.; Ramírez-Verduzco, L.; Hernández, J. Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel. *Fuel* **2015**, *147*, 9–17. [[CrossRef](#)]
7. Mostafaei, M.; Javadikia, H.; Naderloo, L. Modeling the effects of ultrasound power and reactor dimension on the biodiesel production yield: Comparison of prediction abilities between response surface methodology (RSM) and adaptive neuro-fuzzy inference system (ANFIS). *Energy* **2016**, *115*, 626–636. [[CrossRef](#)]
8. Miraboutalebi, S.; Kazemi, P.; Bahrami, P. Fatty Acid Methyl Ester (FAME) composition used for estimation of biodiesel cetane number employing random forest and artificial neural networks: A new approach. *Fuel* **2016**, *166*, 143–151. [[CrossRef](#)]
9. Raman, A.A.A.; MK, A.; Sulaiman, N. Estimation of vegetable oil-based ethyl esters biodiesel densities using artificial neural networks. *J. Appl. Sci. Data* **2008**, *8*, 3005–3011. [[CrossRef](#)]
10. Wali, W.; Hassan, K.; Cullen, J.; Shaw, A.; Al-Shamma'a, A. Real time monitoring and intelligent control for novel advanced microwave biodiesel reactor. *Measurement* **2013**, *46*, 823–839. [[CrossRef](#)]
11. Fayyazi, E.; Ghobadian, B.; Najafi, G.; Hosseinzadeh, B. Genetic algorithm approach to optimize biodiesel production by ultrasonic system. *Chem. Prod. Process Model.* **2014**, *9*, 59–70. [[CrossRef](#)]
12. Sikorski, J.J.; Brownbridge, G.; Garud, S.S.; Mosbach, S.; Karimi, I.A.; Kraft, M. Parameterisation of a biodiesel plant process flow sheet model. *Comput. Chem. Eng.* **2016**, *95*, 108–122. [[CrossRef](#)]
13. Cheng, M.Y.; Prayogo, D.; Ju, Y.H.; Wu, Y.W.; Sutanto, S. Optimizing mixture properties of biodiesel production using genetic algorithm-based evolutionary support vector machine. *Int. J. Green Energy* **2016**, *13*, 1599–1607. [[CrossRef](#)]
14. Ahmad, I.; Kano, M.; Hasebe, S. Dimensions and Analysis of Uncertainty in Industrial Modeling Process. *J. Chem. Eng. Jpn.* **2018**, *51*, 1–11. [[CrossRef](#)]

15. Helton, J.C.; Cooke, R.M.; McKay, M.D.; Saltelli, A. Sensitivity analysis of model output: SAMO 2004. *Reliab. Eng. Syst. Saf.* **2006**, *10*, 1105–1108. [[CrossRef](#)]
16. Refsgaard, J.C.; van der Sluijs, J.P.; Højberg, A.L.; Vanrolleghem, P.A. Uncertainty in the environmental modelling process—A framework and guidance. *Environ. Model. Softw.* **2007**, *22*, 1543–1556. [[CrossRef](#)]
17. Van Der Heijden, K. Scenarios and forecasting: Two perspectives. *Technol. Forecast. Soc. Chang.* **2000**, *65*, 31–36. [[CrossRef](#)]
18. Linkov, I.; Burmistrov, D. Model uncertainty and choices made by modelers: Lessons learned from the international atomic energy agency model intercomparisons. *Risk Anal. Int. J.* **2003**, *23*, 1297–1308. [[CrossRef](#)]
19. Butts, M.B.; Payne, J.T.; Kristensen, M.; Madsen, H. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *J. Hydrol.* **2004**, *298*, 242–266. [[CrossRef](#)]
20. Duan, Q.; Sorooshian, S.; Gupta, V.K. Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.* **1994**, *158*, 265–284. [[CrossRef](#)]
21. Doherty, J.; Johnston, J.M. Methodologies for calibration and predictive analysis of a watershed model 1. *JAWRA J. Am. Water Resour. Assoc.* **2003**, *39*, 251–265. [[CrossRef](#)]
22. Saltelli, A.; Chan, K.; Scott, E.M. *Sensitivity Analysis*; Wiley: New York, NY, USA, 2000; Volume 1.
23. Knothe, G.; Matheaus, A.C.; Ryan, T.W., III. Cetane numbers of branched and straight-chain fatty esters determined in an ignition quality tester. *Fuel* **2003**, *82*, 971–975. [[CrossRef](#)]
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
25. Ahmad, I.; Mabuchi, H.; Kano, M.; Hasebe, S.; Inoue, Y.; Uegaki, H. Data-Based Ground Fault Diagnosis of Power Cable Systems. *SICE J. Control Meas. Syst. Integr.* **2013**, *6*, 290–297. [[CrossRef](#)]
26. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
27. Helton, J.C.; Johnson, J.D.; Sallaberry, C.J.; Storlie, C.B. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliab. Eng. Syst. Saf.* **2006**, *91*, 1175–1209. [[CrossRef](#)]
28. Perez, R.A. Uncertainty Analysis of Computational Fluid Dynamics via Polynomial Chaos. Ph.D. Thesis, Virginia Tech, Blacksburg, VA, USA, 2008.
29. Oladyshkin, S.; Nowak, W. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliab. Eng. Syst. Saf.* **2012**, *106*, 179–190. [[CrossRef](#)]
30. Rajabi, M.M.; Ataie-Ashtiani, B.; Simmons, C.T. Polynomial chaos expansions for uncertainty propagation and moment independent sensitivity analysis of seawater intrusion simulations. *J. Hydrol.* **2015**, *520*, 101–122. [[CrossRef](#)]
31. Sudret, B.; Der Kiureghian, A. *Stochastic Finite Element Methods and Reliability: A State-of-the-Art Report*; Department of Civil and Environmental Engineering, University of California Berkeley: Berkeley, CA, USA, 2000.
32. Ahmad, I.; Ali, G.; Bilal, M.; Hussain, A. Virtual sensing of catalytic naphtha reforming process under uncertain feed conditions. In Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 3–4 March 2018; pp. 1–6.
33. Ahmad, I.; Mabuchi, H.; Kano, M.; Hasebe, S.; Inoue, Y.; Uegaki, H. Data-Based Fault Diagnosis of Power Cable System: Comparative Study of k-NN, ANN, Random Forest, and CART. In Proceedings of the 18th IFAC World Congress, Milano, Italy, 28 August–2 September 2011; pp. 12880–12885.

