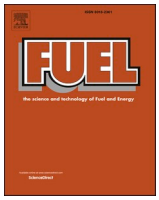
Fuel 332 (2023) 126177

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel



Full Length Article

Fast characterization of biodiesel via a combination of ATR-FTIR and machine learning models

Chao Chen [a], Rui Liang [a], Shaige Xia [b], Donghao Hou [a], Boré Abdoulaye [a], Junyu Tao [b,*], Beibei Yan [a,c], Zhanjun Cheng [a], Guanyi Chen [b,d]

[a] School of Environmental Science and Engineering, Tianjin University, Tianjin 300350, China
[b] School of Mechanical Engineering, Tianjin University of Commerce, Tianjin 300134, China
[c] Tianjin Key Lab of Biomass Wastes Utilization/Tianjin Engineering Research Center of Bio Gas/Oil Technology, Tianjin 300072, China
[d] School of Science, Tibet University, Lhasa 850012, China

ARTICLE INFO

Keywords:
Biodiesel
Fuel properties
ATR-FTIR
Machine learning

ABSTRACT

This study proposed a fast characterization method for biodiesel based on attenuated total reflection flourier-transform infrared spectroscopy and machine learning models. The concerning characteristics of biodiesel include unsaturated group content, O content, and contents of four representative esters. A total of 71 biodiesel samples were produced from a lab-scale reactor. Their spectral data and characteristics were collected and used as training data for machine learning models. The established model framework consisted of two data compression sections, a classification section, and a regression section, all of which use machine learning models, such as principal component analysis, support vector machine, artificial neural network, and random forest. The accuracy, correlation, and sensitivity of the proposed method were evaluated and optimized. Furthermore, the interpretation of the models was discussed. The results showed that the principal component analysis model was a satisfactory preprocessing procedure for the downstream classification and regression models. Under the optimal model parameters, the integrated framework could reach an average accuracy rate of 93.18% and a Pearson correlation coefficient of 0.92. Principal component number 4 and 5 showed the highest sensitivity towards the predicting results, implying that their highly weighted wavenumber ranges and the correlated functional groups played the most important role in the predicting process. The findings of this study could lead to a simple and efficient approach for characterizing the properties of biodiesel, which in turn could promote the development of similar biomass-derived liquid fuels.

1. Introduction

Diesel is a fossil fuel-derived liquid fuel. Its durability is provided by its low viscosity and high density [1]. As a result, the demand for petroleum-based fuels is also increasing fast. However, on the one hand, with the rapid growth of the demand for petroleum-based fuels, the consequent shortage of petroleum energy has become increasingly prominent. On the other hand, considering fossil fuels' unsustainable and carbon footprint emission problems [2], alternative diesel resources are urgently needed. Biodiesel is an environmentally friendly renewable energy source now widely used as an alternative liquid fuel for compression-ignition engine traffic and transportation [3]. Because it produces biodiesel with qualities that are highly similar to diesel fuel, transesterification is the preferred method in the production process of biodiesel worldwide [4]. According to the BP Statistical Review of World Energy 2019 report, biodiesel production reached 34.9 million tons of oil equivalent, of which Europe accounted for 37 % [5]. Moreover, biodiesel, which has good environmental protection and application characteristics [6], is increasingly accepted by the public because of its renewable, bio-degradable, non-toxic, and low-emission advantages [7], and is used in blending with mineral diesel in developed countries around the world [8].

In the actual operation of the biodiesel production industry, waste cooking oil is the main raw material for biodiesel production. In China, the annual output of waste cooking oil has surpassed 90 million tons, posing a threat to the environment [9]. Therefore, the production of biodiesel from waste cooking oil effectively reduces the raw material cost of biodiesel. Some studies have reported the economic evaluation of biodiesel production from edible waste oil and found that it can achieve

* Corresponding author.
E-mail address: taojunyu@tjcu.edu.cn (J. Tao).

https://doi.org/10.1016/j.fuel.2022.126177
Received 22 May 2022; Received in revised form 26 September 2022; Accepted 28 September 2022
Available online 10 October 2022

Nomenclature

ANN	Artificial neural network
ATR-FTIR	Attenuated total reflection flourier-transform infrared spectroscopy
C19:4	9,12-Octadecadienoic acid (Z,Z)-, methyl ester
C19:2	9-Octadecenoic acid (Z)-, methyl ester
C19:1	Methyl stearate
C17:1	Hexadecanoic acid, methyl ester
CFPP	Cold filter plugging point
CP	Cloud Point
GC–MS	Gas Chromatography-Mass Spectrometer
MRE	Mean Relative Error
PCA	Principal component analysis
PCs	Principal components
RF	Random Forest
SVM	Support Vector Machine

the goal of profit. It also proves that installing large-scale biodiesel production facilities may positively impact and create new employment opportunities for the poorest communities [10]. However, many obstacles still hinder the development of biodiesel industry. Some studies used interpretative structural modeling to determine the relationship between obstacles [11]. Besides, the characteristics of the feedstock (e.g., moisture content, different cooking habits, and non-homogeneous nature of kitchen waste oil, etc.) were often unstable, which might lead to unstable products and create uncertainty about carbon emissions. Therefore, when producing biodiesel, it was essential to obtain the characteristics of the produced biodiesel product, which can play a good command in the selection of raw materials and reaction conditions [12,13], and the production of biodiesel [14].

Pure biodiesels' performance was affected by some important fuel properties (density, viscosity, flash point, heating value, critical properties, cloud point, latent heat of vaporization, etc.) [15]. Among the various characteristics of biodiesel, low-temperature mobility is one that has a significant impact on its use. The low-temperature mobility of biodiesel is primarily evaluated by the cloud point (CP) [16] and the cold filter plugging point (CFPP) [17,18]. There has been some previous studies on biodiesel low-temperature performance prediction about CP [19]. However, CP is a static and terminal test that cannot achieve real-time dynamic and accurate detection. Furthermore, the composition and content of each ester in biodiesel were found to be strongly related to CFPP [20]. Therefore, it is necessary to investigate the component content of each ester in the ester mixture of biodiesel components to achieve real-time monitoring and testing of biodiesel quality. There are two critical challenges with using biodiesel in diesel engines: oxidative stability and unsaturated bond concentration [21]. Both of these factors are important indicators of biodiesel performance. Currently, these properties are usually measured by a range of analyzers, such as Gas Chromatography-Mass Spectrometer (GC–MS), oxygen bomb calorimeters, and elemental analyzers for elemental compositions [22,23]. A large number of test procedures are required to obtain all these properties, and although the experimental measurement of these properties is not complicated, it requires considerable cost and time.

Compared with the above analytical and testing techniques, infrared spectroscopy is a potentially promising method for the properties characterization of biodiesel. Still, its application in this field has not been fully explored. The strengths of infrared spectroscopy are that, the spectrogram is very sensitive to the type and content of functional groups in biodiesel and is closely related to its elemental composition, unsaturated bond concentration, and other important properties. However, infrared spectroscopy presents a significant challenge because it requires expert assistance to extract useful information from the spectra. This process is also complex and time-consuming.

As the field of computer science has progressed rapidly, machine learning models are applied and have found widespread use in several data-processing fields. Recent years have seen an expansion of studies utilizing machine learning to characterize and analyze biomass-derived compounds, generally from two aspects. On the one hand, machine learning techniques have a large number of studies and applications in biodiesel characterization and modeling, including cetane number [24], calorific value [25], densities of ternary blends [26], and other applications. On the other hand, there is a large amount of research on machine learning techniques in modeling biodiesel energy utilization, such as combustion parameters [27,28], exhaust gas emissions [29,30], and other applications. Considering the powerful characterization capability of spectroscopy and the wide application of machine learning in biodiesel production, the combination of spectroscopic techniques and machine learning models is likely to be a promising approach to achieving fast and easy characterization of bio-liquid fuels. However, to our knowledge, very few articles have shed light on this research direction.

Based on the above review, the hypothesis of combining spectroscopy with machine learning to achieve fast, easy and non-destructive measurements of biomass liquid fuel performance parameters is proposed. To validate the feasibility of the proposed methodology, a large number of biodiesel samples were collected. Machine learning model framework based on principal component analysis, support vector machine, artificial neural network, and the random forest was constructed to predict the unsaturated group content, O content, and contents of four representative esters in biodiesel. Biodiesel has environmental and energy benefits, and its quality monitoring is an important factor affecting development. Therefore, this method proposed in this work provides a great impetus for the development of biodiesel. The results of this study could inspire new approaches to biodiesel quality control and testing, pave the way for further advancements of liquid biofuels, and provide fundamental knowledge in this field.

2. Experimental and method

2.1. Biodiesel sample preparation

This paper produced biodiesel samples from waste kitchen oil and methanol using a lab-scale transesterification reactor. To ensure the generality of the established method, the reacting conditions for the biodiesel samples were selected as orthogonal as possible. GC–MS was used to characterize the compositions of different biodiesel samples and the infrared spectra by using ATR-FTIR. According to the GC–MS results, the biodiesel samples were mainly composed of esters. C19:4, C19:2, C17:1, and C19:1 were the dominant components of biodiesel esters, accounting for more than 99 % of the total esters. The unsaturated group content and O content were also obtained from the GC–MS results.

In the following data set establishing process, the spectra data were used as input of the model framework, and the biodiesel properties (unsaturated groups content, O content, and contents of the four representative esters) generated from GC–MS were used as the output of the model framework. Each biodiesel sample's input and output parameters were combined to form a sample in the data set. The data set consisting of 71 biodiesel samples were randomly divided into a training set (42 samples), a validation set (14 samples), and a test set (15 samples).

2.2. Attenuated total reflection flourier-transform infrared spectroscopy

ATR-FTIR has the advantages of fast detection and easy operation, making it a popular instrument in recent years. ATR-FTIR obtains the structural information of the organic constituents on the surface of the sample by the reflected signal from the surface of the sample, instead of by the signal through the sample, which simplifies the sample

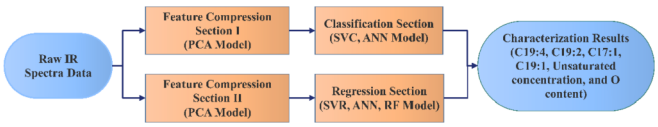


Fig. 1. Scheme of the hybrid predicting model framework.

preparation process and greatly expands the application of infrared spectroscopy. This study conducted ATR-FTIR tests of the model biodiesel using a Thermo Fisher IS50 analyzer. The peak position range was set from 400 to 4000 cm^{-1}, the resolution was set to 0.482 cm^{-1}, and 7469 peak positions were obtained from the collected data. Air background spectra were collected before sample spectra. The working process of ATR-FTIR was that each test sample is automatically measured 32 times, the abnormal value was eliminated at the same time, and the final data result was obtained by averaging 32 measurement data results. Therefore, the reliability and stability of the data in ATR-FTIR were determined by the calibration procedure of the machine itself and continuous cyclic repeated measurements. The average result was used for the following analysis.

2.3. Model establishment

The hybrid prediction model framework used in this study is shown in Fig. 1. Visualization of the gathered spectra data revealed that the presence or absence of a particular ester in a sample might significantly affect the spectral pattern. As a result, a classification component was introduced to improve the model framework's predictive performance.

However, excessive data dimensionality will increase the complexity of the model, especially for small sample data, and the final trained model will have poor generalization. Therefore, feature extraction and compression of the spectra are performed using PCA, which can achieve the effect of optimizing the model. The principal components (PCs) are applied to represent the original data after dimensionality reduction while retaining most of the original information. Each PC is a linear combination of absorbance values at different wavenumbers, as in Eq. (1).

$$PC_m = \sum_{l}^{n} l_{m,n} \cdot a_n \tag{1}$$

PCA, SVM, ANN, and RF were conducted by the Scikit-learn v0.21.2 package in Python 3.7.3 programming environment. And the number of principal components (PC) up to 20 was studied.

The neural network has three processing layers, which are the input layer, the output layer, and the hidden layer. The three layers consist of many interconnected nodes. Each node represents a specific output function, and each connection between two nodes represents a weighted value of the signal passing through that connection, which indicates the strength of the information between these two neurons. It also contains different hyperparameters (activation function, optimization function, number of hidden layers). These parameters are important factors that affect the effectiveness of the neural network model. Previous studies have shown that simple ANN can play a good role in classification and recognition [31,32]. Therefore, referring to the empirical formula, this study sets two hidden layers in ANN with five neural nodes in each layer. The study sets four types in the activation function and four types in the optimizer.

Both SVC and SVR belong to support vector machines, which are considered among the most efficient algorithms for handling small sample data and high dimensionality. The kernel functions take vectors in the original space as input vectors and return the dot product of vectors in the feature space. We studied different kernel functions in support vector machines, including linear kernel functions (linear), polynomial kernel functions(poly), and radial basis kernel functions (rbf). The mathematical definitions of these kernels are shown in Eqs. (2)–(4).

$$linear: k(x,y) = x'y + c \tag{2}$$

$$poly: k(x,y) = (axTy + c)d \tag{3}$$

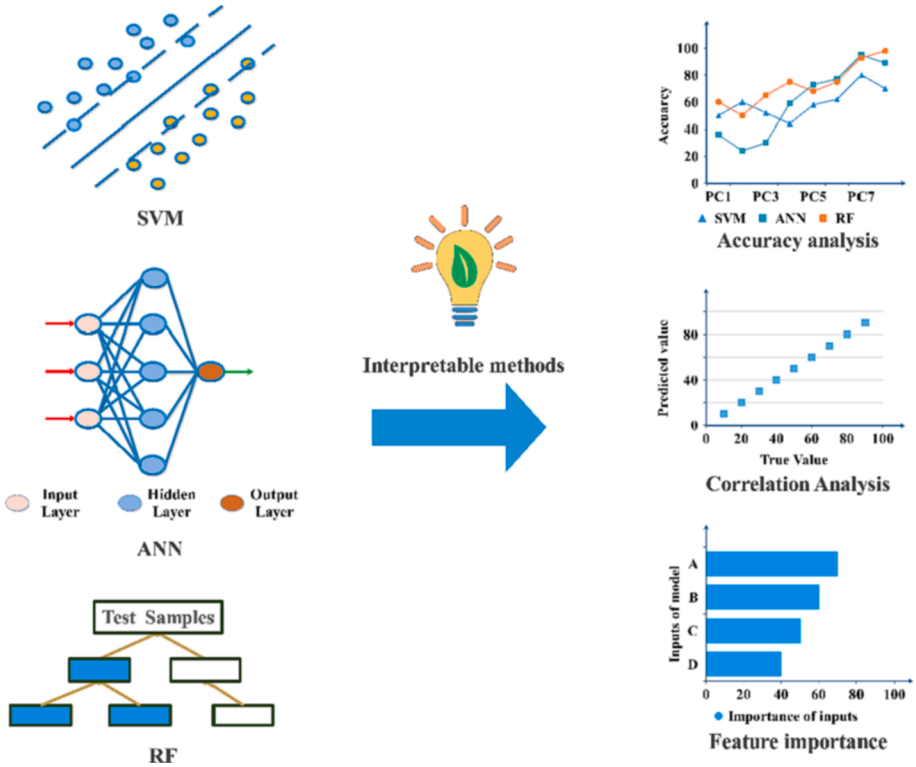


Fig. 2. Graphical introduction of several major machine learning models and interpretable methods.

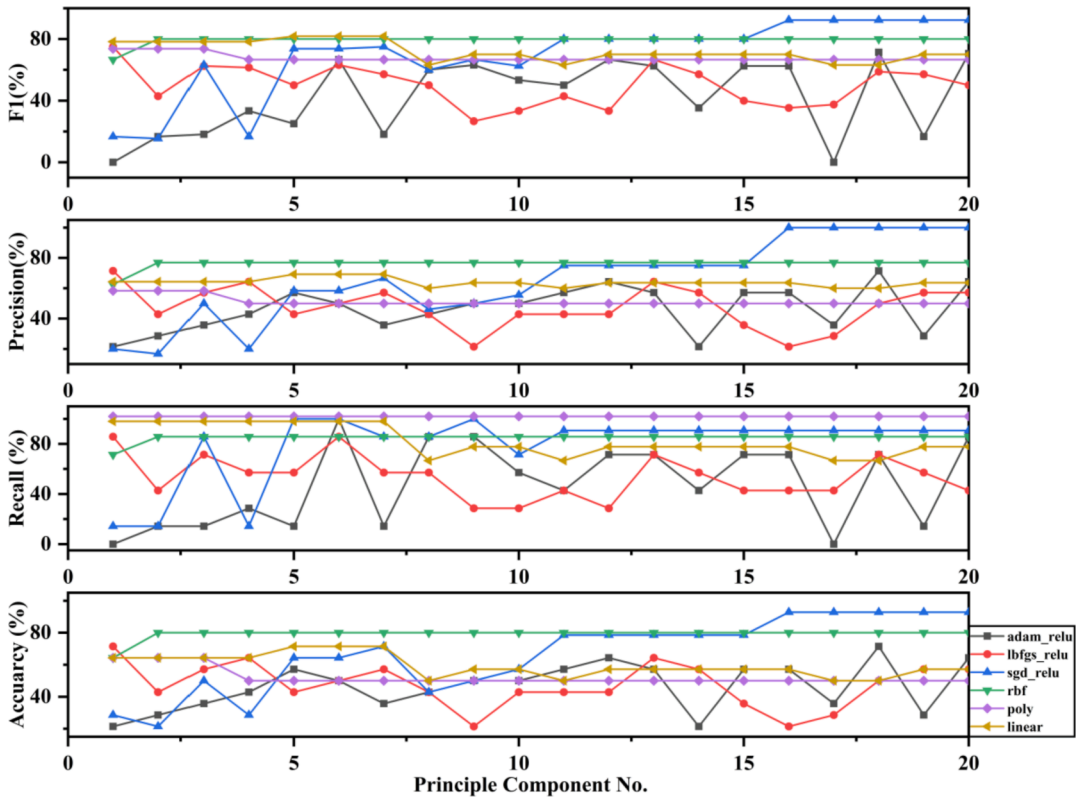


Fig. 3. Effect of the principal component number on the performance of the classification section by (A) C19:4.

$$rbf: k(x,y) = \exp(-\gamma\|x-y\|) \tag{4}$$

where x and y are data vectors; k (x, y) is the kernel for data x and y; then a, c, and d are all the constant parameters in these kernels. In this study, linear, poly, and rbf were studied.

A random forest is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. Under the RF algorithm, this study sets up two optimizers for model prediction concerning the empirical formula. Fig. 2 focuses on the graphical presentation of machine learning models and interpretable methods.

2.4. Data evaluation

The performance of the classification model was evaluated by accuracy, precision, recall, and F1 scores:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{5}$$

where TP is the amount of true positive samples (the value of the actual class is 1 and the value of the predicted class is also 1), TN is the amount of true negative samples (the value of the actual class is 0 and the value of the predicted class is also 0), FP is the amount of false-positive samples (the value of the actual class is 0 but the value of the predicted class is 1), FN is the amount of false-negative samples (the value of the actual class is 1 but the value of the predicted class is 0). The performance of the regression model was evaluated using the average accuracy rate:

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{7}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{8}$$

The performance of the regression model was evaluated by mean relative error (MRE):

$$MRE = \frac{\frac{1}{n} \times \sum_{i=1}^{n} |x_i - y_i|}{|\overline{y}|} \times 100\% \tag{9}$$

where n is the total number of investigated samples, x_i is the predicted value of sample i, y_i is the actual value of sample i, and y is the mean of the actual values of all evaluated samples.

As mentioned above, the spectral data were dimensionally reduced into a series of PCs. According to the loadings of each PC, the amount of spectral information each PC contained could be observed. However, when a specific PC has more information, it doesn't mean that the PC plays a more important role in the predicting model. A sensitivity analysis was conducted to understand the importance and role of each PC. During sensitivity analysis, each PC was changed by ± 30 %, and the changes in the predicting results were evaluated. A higher change in the predicting results is recognized to have a higher sensitivity of the specific PC, and vice versa. The changes in the predicting results were quantified by disturbance ratio:

$$\eta_{1.3PC} = \frac{\sum_{i=1}^{20}(H_{1.3PCi} - H_{1.0PCi})}{\sum_{i=1}^{20} H_{1.0PCi}} \tag{10}$$

$$\eta_{0.7PC} = \frac{\sum_{i=1}^{20}(H_{0.7PCi} - H_{1.0PCi})}{\sum_{i=1}^{20} H_{1.0PCi}} \tag{11}$$

where $\eta_{1.3PC}$ is the value of PC disturbance ratio at + 30 %, $\eta_{0.7PC}$ is the value of PC disturbance ratio at − 30 %, $H_{1.3PCi}$ is the accuracy of 130 % PC number from 1 to 20, $H_{0.7PCi}$ is the accuracy of 70 % PC number from 1 to 20, and $H_{1.0PCi}$ is the accuracy of the original PC number from 1 to 20.

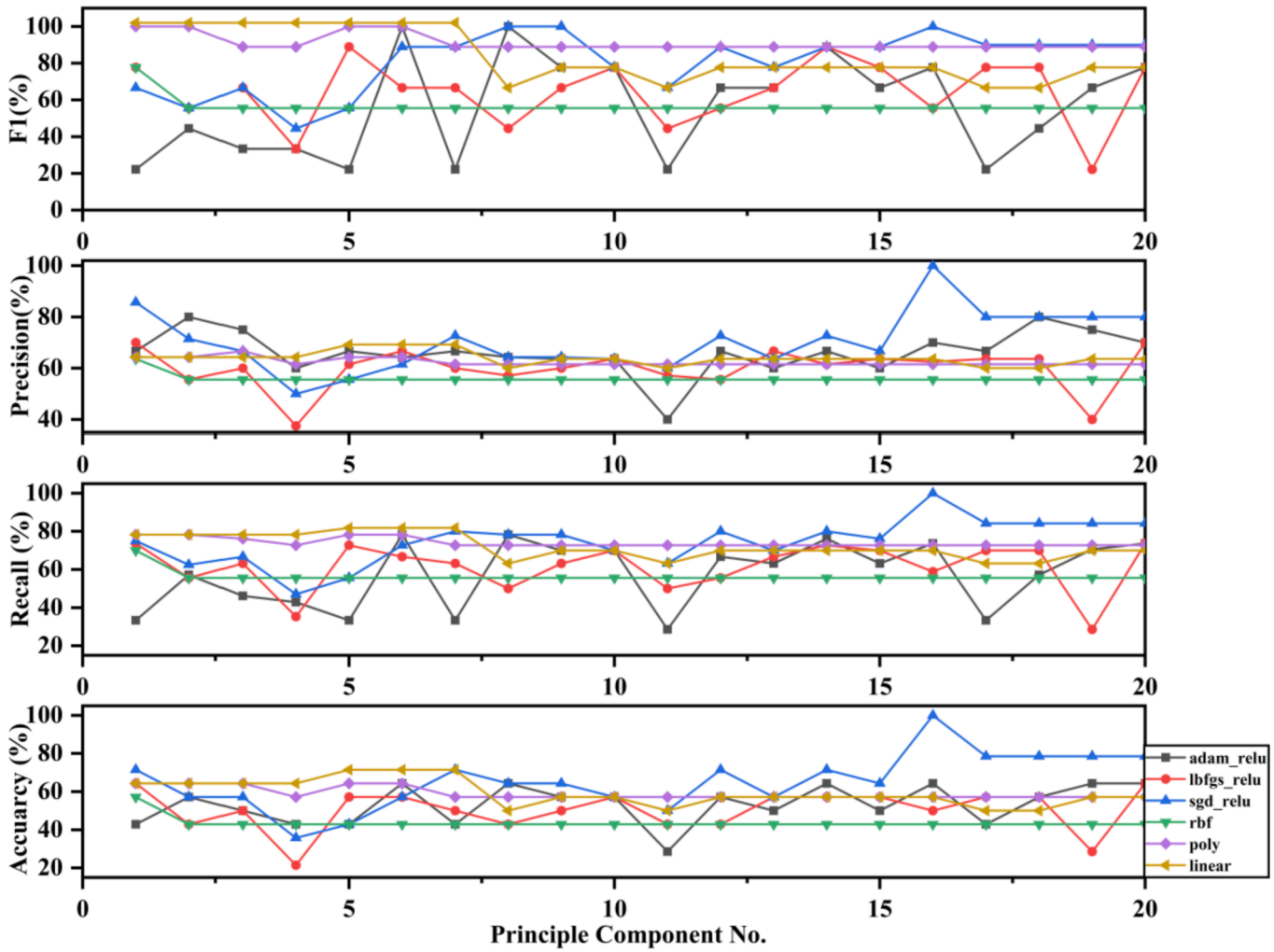


Fig. 4. Effect of the principal component number on the performance of the classification section by (B) C19:2.

3. Results and discussion

3.1. Evaluation and optimization of the classification section

A classification section was applied and evaluated with different PC numbers in the feature compression process of the IR spectrum, and the biodiesel components of C19:4, C19:2, C17:1, and C19:1 were classified. The optimal results of the activation functions of adam, lbfgs, and sgd in ANN were extracted in the classification stage (see Supplementary data), and the results of the three kernel functions of linear, poly, rbf in SVC were also extracted. Higher accuracy, precision, or recall represented better classification performance. Precision and recall are mutually restrictive. F1-score comprehensively considered precision and recall. The closer F1-score is to 1, the better the classifier performance will be. The classification results with different functions, activation functions, and different numbers of PCs are shown in Figs. 3–6.

In the biodiesel component C19:4, the models with linear and poly were easier to classify when the PC number was low, as shown by the recall results. However, many samples were also misclassified under this condition. From the accuracy, precision, and F1 score results, it was easy to find that the model achieved the best classification with PC number reaching 16 under the optimizer of ANN-sgd-relu. Similarly, in the biodiesel component C19:2, it was found that when the PC number reached 16 under the optimizer of ANN-sgd- relu, the evaluations of accuracy, precision, recall, and F1 score all reached a stable state of 100 %, suggesting all samples were correctly classified. In the same way, the model also performed well for the biodiesel C19:1, with the PC number reaching 16 under ANN-lbfgs-relu.

For the biodiesel component C17:1, the best classification was achieved when the PC number reached 13. However, it did not show a stable state with an increasing PC number. This may be due to the factor of randomness, which does not have stability. Fig. 5 revealed that the classification effect of the model reached a stable state after PC number 16.

As can be seen from Figs. 3–6, the classifier effect gradually reached a steady state as the PC increased. The optimal result of the classification is summarized as follows in Table 1. Regarding the optimal results, ANN outperformed SVC in general. Furthermore, the classification effect reached the optimal state when PC numbers reached 16 or more.

3.2. Evaluation and optimization of the regression section

Regression models were used to predict the results of biodiesel components C19:4, C19:2, C17:1, and C19:1, unsaturated concentration, and O content. The regression model was performed using the training data set. Then the cross-validation set was used to determine the optimal model parameters. Finally, the optimal parameters were brought into the model to verify the generalization ability against test data. The MRE was used to evaluate the performance of the regression prediction, and it reflected the difference between the predicted and actual results. The smaller the MRE is, the better the regression model is. The regression results for different models and different PC numbers are shown in

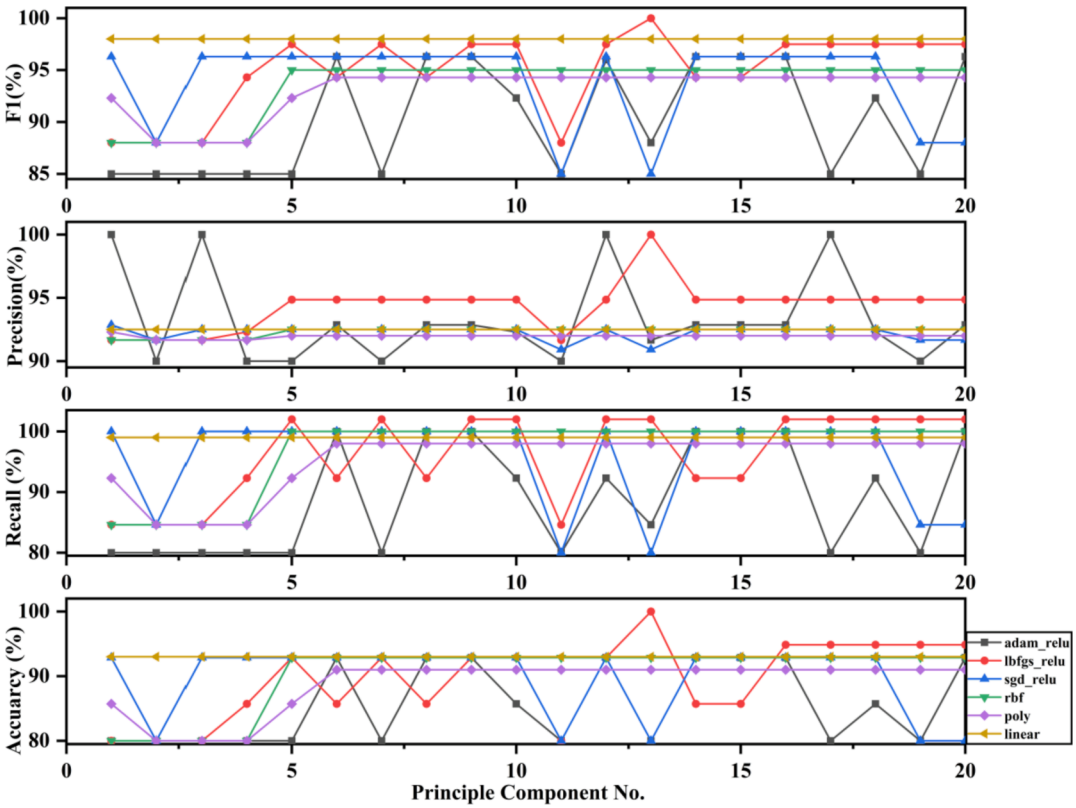


Fig. 5. Effect of the principal component number on the performance of the classification section by (C C17:1.

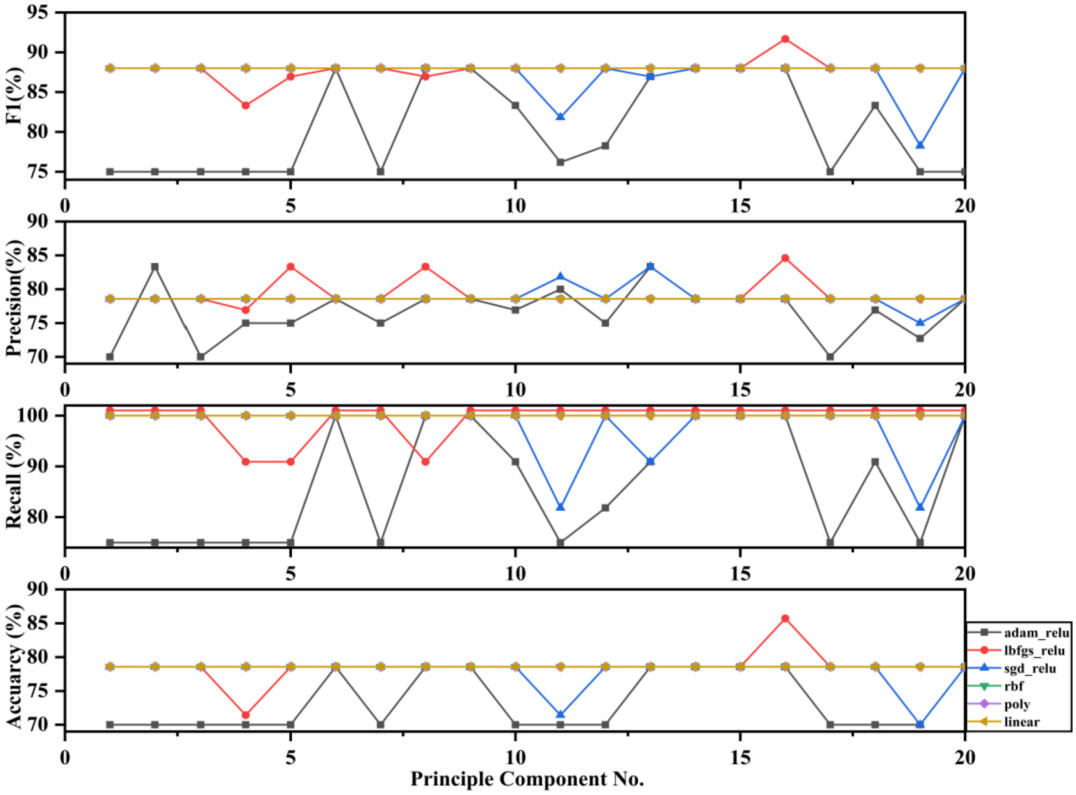


Fig. 6. Effect of the principal component number on the performance of the classification section by (A) C19:1.

Table 1
Summary of accuracy, precision, recall, and F1 for different prediction objectives with the optimal number of PCs and models.

	PC number	Model	Accuracy	Precision	Recall	F1
C19:4	16	ANN-sgd-relu	92.9 %	100 %	85.7 %	92.3 %
C19:2	16	ANN-sgd-relu	100 %	100 %	100 %	100 %
C17:1	16	ANN-lbfgs-relu	92.9 %	92.9 %	100 %	96.3 %
C19:1	16	ANN-lbfgs-relu	85.7 %	84.6 %	100 %	91.7 %

Fig. 7.

With an increasing PC number, the regression model obtains more information from the original IR spectrum, and its prediction grows closer to the actual state. In general, the MRE of SVR using a linear kernel, RF, and ANN decreased when PC was less than 4 or 5, meaning that the top 4 or 5 PCs contain the most quantitative data. The information retrieved from the IR spectrum by the generated regression model improves as the number of PCs increases.

Contrary to ANN and RF models, SVR using rbf saw no significant change in MRE with increasing PC number. For the biodiesel components of C19:2 and C17:1, the MRE results gradually increased after the PC number 4 or 5, indicating that the increase of PC number led to a more complex model structure, which exacerbated the overfitting problem during model training. The optimal parameters for the biodiesel components of C19:2, C17:1, and unsaturated concentration were ANN-lbfgs-identity, RF-sgd, and RF-sgd, respectively. For the biodiesel components of C19:1O content, the optimal parameters were linear and rbf in SVR, and the optimal MRE results were 5.8 % and 8.9 %, respectively. Furthermore, ANN-lbfgs-tanh was the optimum parameter for the C19:4 biodiesel components, yielding an MRE of 6.9 %. In other words, the MRE for the six optimal parameters ranged from 3.7 % to 8.9 %. This result might not be as accurate as traditional characterization methods (e.g., using GC–MS, elemental analyzers, and calorimeters), but the measurement process was significantly faster.

Table. 2 summarizes the optimal parameters in the validation and

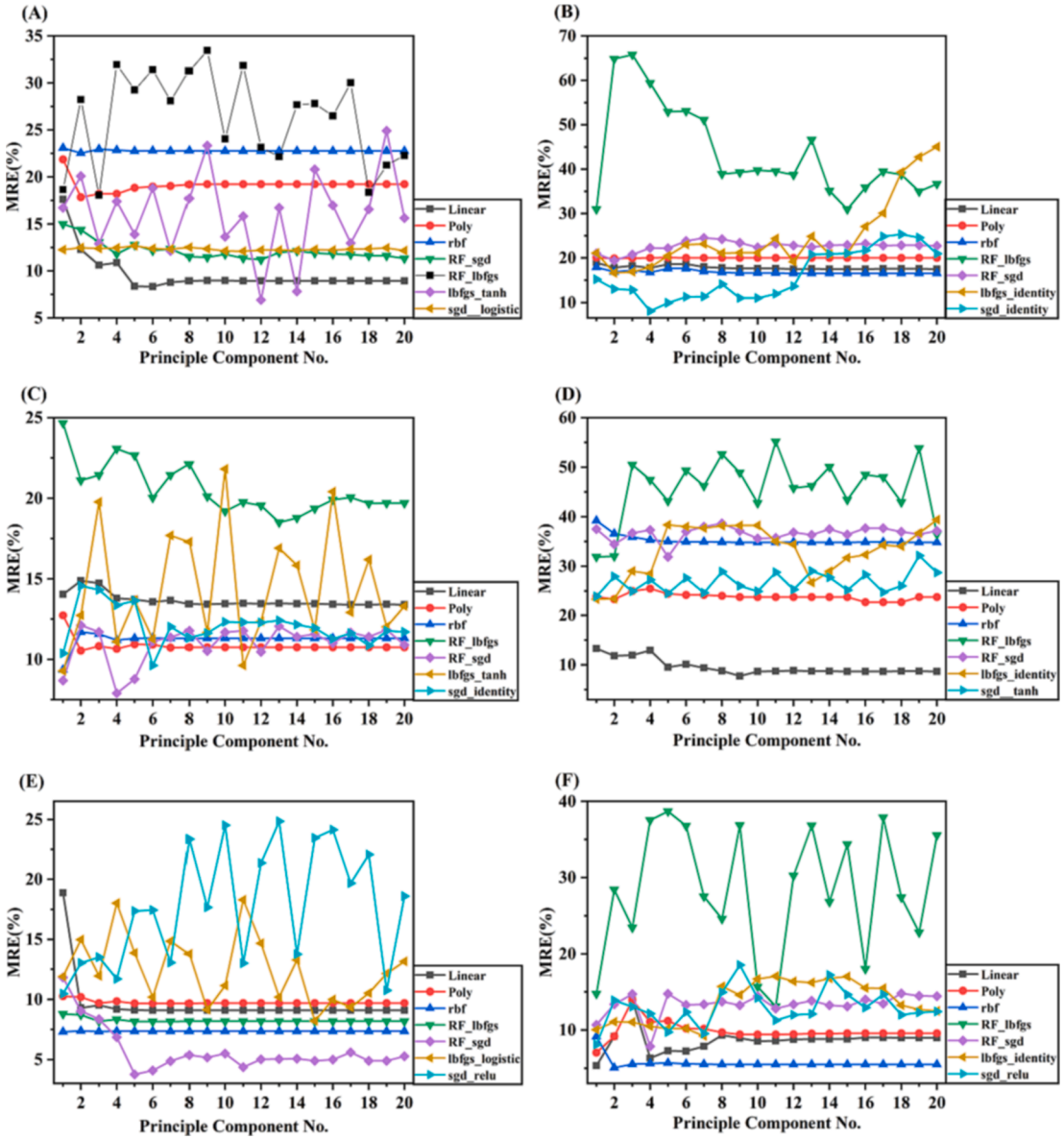


Fig. 7. Effect of the principal component number on the performance of the regression section for predicting A) C19:4, B) C19:2, C) C17:1, D) C19:1, E) unsaturated concentration, F) O content.

Table 2
Summary of the accuracies of different predicting targets in cross-validation and test set under the optimal PC number and kernel function.

	Cross-validation set	Test set	PC number	PCC	Kernel function
C19:4	93.08 %	93.82 %	12	0.97	ANN-lbfgs-tanh
C19:2	91.92 %	90.6 %	4	0.88	ANN-lbfgs-identity
C17:1	91.11 %	91.15 %	4	0.90	RF-sgd
C19:1	91.37 %	92.86 %	9	0.95	Linear
Unsaturated concentration	96.25 %	97.03 %	5	0.94	RF-sgd
O content	94.20 %	93.59 %	7	0.90	Rbf

test set. The MRE for C19:4, C19:2, C17:4, and C19:1 reached 6.18 %, 9.40 %, 8.85 %, and 7.14 %, respectively, in the test set. The MRE of unsaturated concentration and O content reached 2.97 %, and 6.41 %, respectively. In general, a model is considered "good enough" if its cross-validation accuracy closely matches that of the test set. Table. 2 verified that the established model had excellent generalization ability.

However, the low MRE of the cross-validation set did not mean that the predicted values had a very close relationship with the actual values. For example, when the span between the actual values of the cross-validation set was small, the prediction results could achieve accurate results, which might be caused by the relatively dense distribution of the actual values. When the span between the actual values of the cross-validation set became large, the MRE would increase dramatically. Therefore, it was necessary to study the correlation between the set's true and predicted values. To visualize the results of the data more linearly, we log the results of the actual and predicted values. The specific results are shown in Fig. 8 as well as the correlation of each predictor in Table 2.

According to the correlation analysis, the correlation coefficients of the average predicting targets, consisting of C19:4, C19:2, C17:1, C19:1, unsaturated concentration, and O content, were found to be 0.92, which implies a robust correlation. This further demonstrates the stability of the prediction system.

3.3. Interpretation of different PCs in PCA preprocessing

PCA is used for feature extraction and dimensionality reduction. The importance of each PC is determined by the explained variance ratio (EVR). The higher the EVR, the more information the PC reflects about the original IR spectrum. The sum of EVR of all PCs is equal to or less than 1. As mentioned in Section 3.2, the optimal accuracy in regression prediction was achieved with the number of PCs at 4, 5, 7, 8, and 12. Therefore, this study focuses on PC1 to 12 as shown in Fig. 9.

Fig. 9 shows that the EVR of PC1 accounted for up to 73.61 %, indicating that PC1 extracts most of the information in the IR spectrum.

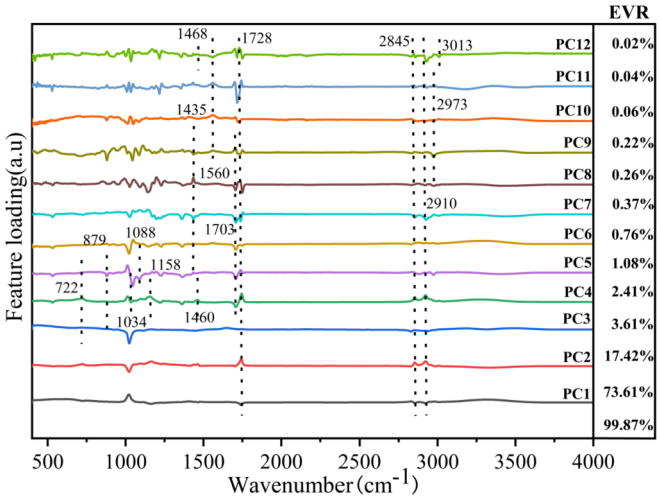


Fig. 9. Feature loadings of top 12 PCs in different wavenumbers.

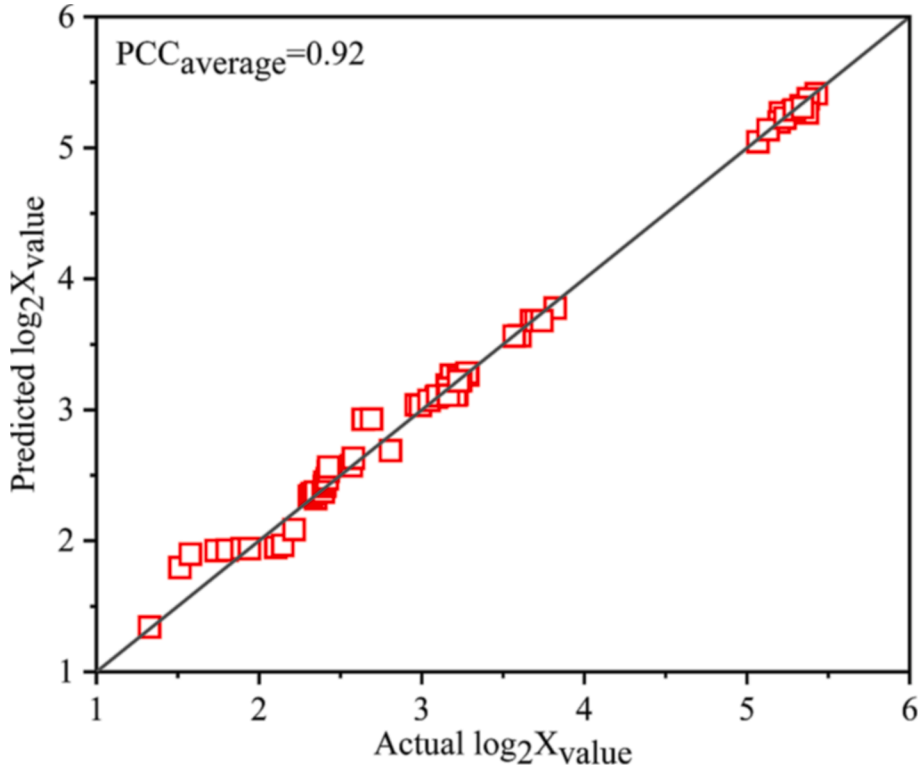


Fig. 8. The scatterplot of the predicted vs the experimental values of log2X: C19:4, C19:2, C17:1, C19:1, unsaturated concentration, and O content.

Table 3
Feature loadings of top 12 PCs in different wavenumbers.

Wavenumber range (cm^{-1})	Functional group
Around 700–1200	-C—C- Alkanes
Around 1400–1600	Asymmetric $–CH_2$, CH_3, O═O
Around 1703	-C—O—O- Ester-based
Around 1728	-C—O—O- Ester-based
Around 2845–2973	-C—H Alkanes
Around 3013	=C—H Olefins

More importantly, it was found that PC1 ~ PC12 could reach 99.97 %, which means that most of the information of the IR spectrum can be retained from PC1 to PC12. Through observation and comparison, particular peaks between PC4, PC5, PC8, and PC12 were identified, and these peaks affected the accurate prediction of the machine learning model algorithm. Further in-depth analysis revealed that these particular peak locations and the model optimal PC number achieved mutual corroboration. Alkanes were determined to be the most common functional group in the IR spectra data, with the majority of characteristic wavenumber bands on PC4 and PC5 located between 700 cm-1 and 1200 cm^{-1}. The particular peaks on PC8 are at wavenumbers 1435 cm^{-1}, 1468 cm^{-1}, and 1560 cm^{-1}, respectively (see Table 3). Additionally, by examining the relationship between the characteristic peak positions and the functional groups, $–CH_2$, CH_3, and O═O bonds, respectively, were found to influence the predicted accuracy of the samples in the machine learning model algorithm. Meanwhile, PC8 showed characteristic peaks at 1703 and 1728 cm^{-1}; corresponding to the ester functional group shared by C19:4, C19:2, C17:4, and C19:1. Although these PC numbers only had low EVR values and contain few pieces of information, they might contain important quantitative information extracted from the IR spectra data, hence paving the way for future studies.

3.4. Sensitivity analysis of different PCs

As mentioned above, the PC containing greater information does not necessarily mean that the PC is important. To investigate the effect on the model accuracy with different numbers of PCs, we also investigated the significant factor of each PC, which is also called the weighting factor. The impact on the reliability of the model's predictions is expressed as a multiplier, which can either increase or decrease the significance of the influencing variables. In this study, the optimal number of PCs obtained in Section 3.2 was changed by ± 30 % for each PC during the sensitivity analysis. The results are shown in Fig. 10.

From the results shown in Fig. 10, C19:4 causes a significant perturbation in PC1 ~ 20. This confirms the results of the optimal parameters in C19:4: ANN-lbfgs-tanh, where the slope of PCA between 1 and 20 in Section 3.2 has a large fluctuation. Similarly, the biodiesel

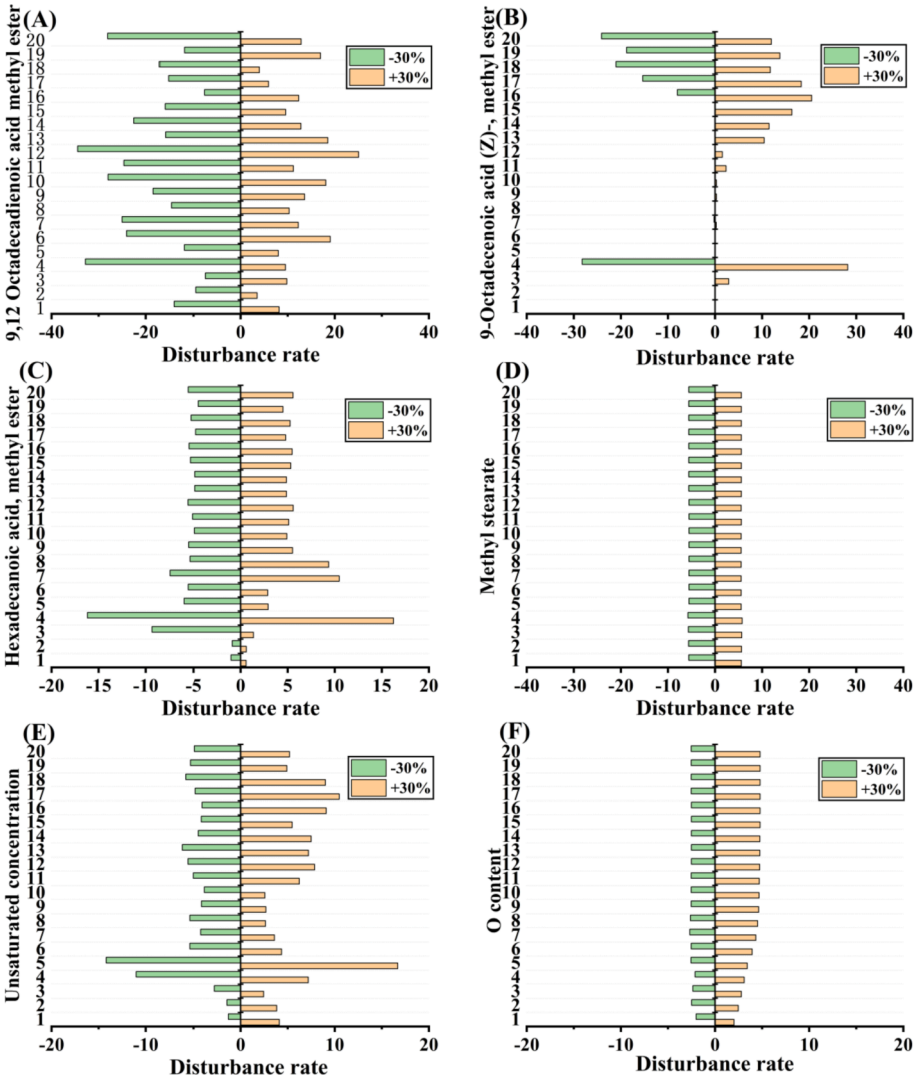


Fig. 10. Sensitivity analysis for predicting A) C19:4, B) C19:2, C) C17:1, D) C19:1, E) unsaturated concentration, F) O content.

components of C19:2 and C17:1 were more prominent in their perturbation rates at PC4. This conforms to the results of the optimal parameters in the biodiesel component of C19:2 and C17:1: ANN-lbfgs-identity, RF-sgd, in Section 3.2, where the slope of PCA suddenly shifted from 3 to 4. This is depicted in Fig. 7. The overall perturbation of the biodiesel component C19:1, O content in PC1-20 was uniform and insignificant. This agrees with the results of the optimal parameters in the biodiesel component C19:4: linear, rbf according to Fig. 7, where there were relatively flat. The unsaturated concentration is more prominent at the perturbation value of PC5 in Fig. 10, and this confirms the results of the optimal parameters in the unsaturated concentration: RF-sgd, where there was a sudden change in the slope of PCA from 4 to 5. Overall, the study of the perturbation rates revealed consistency with the prediction of the change in the model from PCA1 to 20 in Section 3.2, which also further validates the reliability of the sensitivity analysis.

4. Conclusion

This paper proposed a fast characterization method for biodiesel via combination of ATR-FTIR and machine learning models. The established model framework's performance was evaluated and optimized, and its working mechanism was comprehensively discussed. The results showed that PCA played an important role in reducing the dimension of the spectral data and thus enhancing the predicting accuracy. The ANN model was more suitable as a classification section than SVM and RF models. In the regression section, RF was more appropriate for C17:1 and unsaturated concentration. SVM was more applicable for C19:1 and O content. Finally, ANN was more suitable in C19:4 and C19:2. Under the optimal model parameters, the predicting accuracy towards the contents of C19:4, C19:2, C17:1, C19:1, unsaturated group, and O element reached 93.82 %, 90.6 %, 91.15 %, 92.86 %, 97.03 %, and 93.59 %, respectively. The accuracy of this paper is close to the prediction results of other related fields, which combine spectral and machine learning. Furthermore, the optimal correlation coefficients of the average predicting targets were found to be 0.92, which implied an extremely strong correlation. Although PC4 and PC5 did not contain much spectral information, they played the most critical role in predicting results and thus should be paid more attention in further studies. This study showed promising potential for the quality control of industrial biodiesel production, which might substantially promote the development and application of liquid biofuels.

CRediT authorship contribution statement

Chao Chen: Conceptualization, Writing – original draft, Methodology, Investigation, Visualization. **Rui Liang:** Validation, Investigation. **Shaige Xia:** Formal analysis. **Donghao Hou:** Validation, Investigation. **Boré Abdoulaye:** Writing – review & editing. **Junyu Tao:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Beibei Yan:** Funding acquisition, Resources, Methodology, Supervision. **Zhanjun Cheng:** Funding acquisition, Resources, Methodology, Supervision. **Guanyi Chen:** Conceptualization, Funding acquisition, Resources, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by the National Key Research and Development Plan of China (2019YFB154003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fuel.2022.126177.

References

[1] Gulum M, Bilgin A. Measurement and prediction of density and viscosity of different diesel-vegetable oil binary blends. Environmental and Climate Technologies 2019;23:214–28.
[2] Rajaeifar MA, Tabatabaei M, Aghbashlo M, Nizami A-S, Heidrich O. Emissions from urban bus fleets running on biodiesel blends under real-world operating conditions: implications for designing future case studies. Renew Sustain Energy Rev 2019;111:276–92.
[3] Mostafaei M. Prediction of biodiesel fuel properties from its fatty acids composition using ANFIS approach. Fuel 2018;229:227–34.
[4] Bilgin A, Gulum M. Effects of various transesterification parameters on the some fuel properties of hazelnut oil methyl ester. Energy Procedia 2018;147:54–62.
[5] Cui Z, Huang S, Wang M, Nie K, Fang Y, Tan T. Improving the CFPP property of biodiesel via composition design: An intelligent raw material selection strategy based on different machine learning algorithms. Renewable Energy 2021;170:354–63.
[6] Kowthaman CN, Senthil Kumar P, Arul Mozhi Selvan V, Ganesh D. A comprehensive insight from microalgae production process to characterization of biofuel for the sustainable energy. Fuel 2022;310.
[7] Hsiao M-C, Chang L-W, Hou S-S. Study of Solid Calcium Diglyceroxide for Biodiesel Production from Waste Cooking Oil Using a High Speed Homogenizer. Energies 2019;12.
[8] Grabowski P, Jarosiński P. Examination of Selected Physicochemical Properties of Biodiesel after Electron Beam Sterilization in Flow System. Energies 2021;14.
[9] Ling YS, Wang HX, Fei XH, Huang T, Shan Q, Hei DQ, et al. Enhancement effect of gamma-irradiation pre-treatment on anaerobic digestion performance of kitchen wastewater. J Cleaner Prod 2022;330.
[10] Avinash A, Murugesan A. Economic analysis of biodiesel production from waste cooking oil. Energy Sources Part B 2017;12:890–4.
[11] Avinash A, Sasikumar P, Murugesan A. Understanding the interaction among the barriers of biodiesel production from waste cooking oil in India- an interpretive structural modeling approach. Renewable Energy 2018;127:678–84.
[12] Cheng F, Bayat H, Jena U, Brewer CE. Impact of feedstock composition on pyrolysis of low-cost, protein- and lignin-rich biomass: a review. J Anal Appl Pyrol 2020;147.
[13] Baloch HA, Nizamuddin S, Siddiqui MTH, Riaz S, Jatoi AS, Dumbre DK, et al. Recent advances in production and upgrading of bio-oil from biomass: A critical overview. J Environ Chem Eng 2018;6:5101–18.
[14] Zhou J, Liu S, Zhou N, Fan L, Zhang Y, Peng P, et al. Development and application of a continuous fast microwave pyrolysis system for sewage sludge utilization. Bioresour Technol 2018;256:295–301.
[15] Gülüm M, Bilgin A. Two-dimensional surface models to predict the density of biodiesel-diesel-alcohol ternary blends. Energy Sources Part A 2019;43:517–87.
[16] Yadav M, Sharma YC. Transesterification of used vegetable oil using BaAl2O4 spinel as heterogeneous base catalyst. Energy Convers Manage 2019;198.
[17] Ziyad BA, Yousfi M, Vander Heyden Y. Effects of growing region and maturity stages on oil yield, fatty acid profile and tocopherols of Pistacia atlantica Desf. fruit and their implications on resulting biodiesel. Renewable Energy 2022;181:167–81.
[18] Anwar M, Rasul MG, Ashwath N. The efficacy of multiple-criteria design matrix for biodiesel feedstock selection. Energy Convers Manage 2019;198.
[19] Mallah TA, Sahito AR. Optimization of castor and neem biodiesel blends and development of empirical models to predicts its characteristics. Fuel 2020;262.
[20] A.P. Azaria, S.A. Bethari, M. Nasikin, Iop. The use of non-ionic surfactants with different Hydrophilic-Lipophilic Balance (HLB) and their effect on flow properties in palm oil biodiesel. 5TH ANNUAL Applied Science and Engineering Conference (AASEC 2020)2021.
[21] Razavi R, Bemani A, Baghban A, Mohammadi AH, Habibzadeh S. An insight into the estimation of fatty acid methyl ester based biodiesel properties using a LSSVM model. Fuel 2019;243:133–41.
[22] Liu C, Wufuer A, Kong L, Wang Y, Dai L. Organic solvent extraction-assisted catalytic hydrothermal liquefaction of algae to bio-oil. RSC Adv 2018;8:31717–24.
[23] Sajjad Ahmad M, Liu H, Alhumade H, Hussain Tahir M, Çakman G, Yıldız A, et al. A modified DAEM: To study the bioenergy potential of invasive Staghorn Sumac through pyrolysis, ANN, TGA, kinetic modeling, FTIR and GC–MS analysis. Energy Convers Manage 2020.;221.
[24] Bemani A, Xiong Q, Baghban A, Habibzadeh S, Mohammadi AH, Doranehgard MH. Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. Renewable Energy 2020;150:924–34.
[25] Bukkarapu KR, Krishnasamy A. A study on the effects of compositional variations of biodiesel fuel on its physiochemical properties. Biofuels 2018;12:523–35.

[26] Gülüm M, Onay FK, Bilgin A. Measurement and estimation of densities of different biodiesel–diesel–alcohol ternary blends. Environ Prog Sustainable Energy 2019;38.
[27] Rajak U, Nashine P, Verma TN, Pugazhendhi A. Performance, combustion and emission analysis of microalgae Spirulina in a common rail direct injection diesel engine. Fuel 2019;255.
[28] Vignesh R, Ashok B. Deep neural network model-based global calibration scheme for split injection control map to enhance the characteristics of biofuel powered engine. Energy Convers Manage 2021;249.
[29] Balamurugan T, Arun A, Sathishkumar GB. Biodiesel derived from corn oil – A fuel substitute for diesel. Renew Sustain Energy Rev 2018;94:772–8.
[30] Esonye C, Onukwuli OD, Ofoefule AU, Ogah EO. Multi-input multi-output (MIMO) ANN and Nelder-Mead's simplex based modeling of engine performance and combustion emission characteristics of biodiesel-diesel blend in CI diesel engine. Appl Therm Eng 2019;151:100–14.
[31] Aniza R, Chen WH, Yang FC, Pugazhendh A, Singh Y. Integrating Taguchi method and artificial neural network for predicting and maximizing biofuel production via torrefaction and pyrolysis. Bioresour Technol 2022;343:126140.
[32] Rashid T, Ali Ammar Taqvi S, Sher F, Rubab S, Thanabalan M, Bilal M, et al. Enhanced lignin extraction and optimisation from oil palm biomass using neural network modelling. Fuel 2021;293.