IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A *bootstrapping soft shrinkage approach and interval random variables selection* hybrid model for variable selection in near-infrared spectroscopy

HASAN ALI GAMAL AL-KAF[1], NAYEF ABDULWAHAB MOHAMMED ALDUAIS[2], ABDUL-MALIK H. Y. SAAD*,[3], KIM SENG CHIA[4], ABDULQADER M. MOHSEN[5], HITHAM ALHUSSIAN[6], AMMAR ABDO MOHAMMED HAIDAR MAHDI[5,7], WAN SAIFUL ISLAM WAN SALAM[7]

[1]Dept. of Computer and information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Malaysia
[2]Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Malaysia
[3]School of Electrical and Electronic Engineering, Universiti Sains Malaysia (USM), Nibong Tebal 14300, Penang, Malaysia
[4]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Malaysia
[5]Dept. of Computer Science, University of Science and Technology, Sana'a, Yemen
[6]Center for Research in Data Science (CERDAS), Institute of Autonomous Systems (IAS), Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Malaysia
[7]Faculty of Mechanical & Manufacturing Engineering, Universiti Tun Hussein Onn Malaysia, 86400, Parit Raja, Johor, Malaysia

Corresponding author: Abdul-Malik H. Y. Saad (e-mail: abdulmalik@usm.my; eng.abdulmalik@gmail.com).

The authors are grateful to Universiti Sains Malaysia (USM) for financially supporting this research.

ABSTRACT High dimensionality problem in spectra datasets is a significant challenge to researchers and requires the design of effective methods that can extract the optimal variable subset that can improve the accuracy of predictions or classifications. In this study, a hybrid variable selection method, based on the incremental number of variables using bootstrapping soft shrinkage method (BOSS) and interval random variable selection (IRVS) method is proposed and named BOSS-IRVS. The BOSS method is used to determine the informative intervals, while the IRVS method is used to search for informative variables in the informative interval determined by BOSS method. The proposed BOSS-IRVS method was tested using seven different public accessible near-infrared (NIR) spectroscopic datasets of corn, diesel fuel, soy, wheat protein, and hemoglobin types. The performance of the proposed method was compared with that of two outstanding variable selection methods i.e. BOSS and hybrid variable selection strategy based on continuous shrinkage of variable space (VCPA-IRIV). The experimental results showed clearly that the proposed method BOSS-IRVS outperforms VCPA-IRIV and BOSS methods in all tested datasets and improved the percentage of the prediction accuracy, by 15.4 and 15.3 for corn moisture,13.4 and 49.8 for corn oil, 41.5 and 50.6 for corn protein, 12.6 and 5.6 for soy moisture, 0.6 and 6.3 for total diesel fuel, 19.9 and 14.3 for wheat protein, and 5.8 and 20.3 for hemoglobin.

INDEX TERMS Hybrid variable selection, model population analysis, weighted bootstrap sampling, partial least squares, and near infrared spectroscopy.

I. INTRODUCTION

In recent years, near-infrared (NIR) spectroscopy has gained wide acceptance in different fields such as agriculture and the petrochemical and pharmaceutical industries by virtue of its advantages in recording spectra for solid and liquid samples. NIR spectra typically consist of broad, weak, non-specific, and overlapped bands and some irrelevant variables [1]. These unrelated variables could lead to wrong or inefficient prediction results. To overcome this problem, a process of multivariate analysis for NIR spectroscopy should be followed as shown in Figure 1. The first step is to have NIR samples as $\boldsymbol{X}$ and the properties of interest as y. Then a pre-processing technique is used to remove physical phenomena in the spectra [2]. Next, the important variables are extracted using a variable section method. Finally, a multivariate calibration model is used to build the relationship between the selected variables and the properties of interest to predict the values of the interesting properties. Variable selection is a critical step in multivariate calibration of NIR spectroscopy. This is because the variable selection step reduces the curse of dimensionality, which results in speeding-up the operating model, providing a better interpretation of a model by selecting the informative variables, and improving the prediction performance by eliminating uninformative variables [3].

Deng et al. proposed a new and effective single variable selection method named BOSS [4]. This method showed a significant improvement of prediction accuracy on three NIR spectroscopic datasets and outperforms partial least square (PLS), Monte Carlo uninformative variable elimination (MCUVE), competitive adaptive reweighted sampling (CARS) and genetic algorithm coupled with partial least square (GA-PLS). The advantages of the BOSS method can be summarized in three aspects—first, the use of soft shrinkage, which lowers the risk of eliminating essential variables. Second, a fair comparison of variables compensates for the influence of collinearity on the regression coefficients because of the use of weighted bootstrap sampling (WBS). Third, the use of model population analysis (MPA), which extracts the information from a large population of sub-models instead of one model to obtain more reliable results by considering the combined effects among variables [4]. Despite these advantages, the BOSS has drawbacks which can be summed up into three aspects as well. First, the BOSS ignores the high correlation among consecutive variables. Second and due to the use of bootstrap sampling that is inappropriate for the dependent data, the BOSS selects fewer variables, which causes missing some informative wavelengths. Third, it cannot avoid over-fitting problem BOSS uses RC, which is susceptible to noises [5], [6].

Most recently, three different methods have been developed and out-performed BOSS method. The first method is a modification of the bootstrapping soft shrinkage approach named new computational method stabilized bootstrapping soft shrinkage approach (SBOSS) [5], in which variables are selected by the index of stability of regression coefficients instead of regression coefficients absolute value. Second, fisher optimal subspace shrinkage (FOSS) [6] that splits variables into some intervals by the information from regression coefficients PLS model, then the weighted block bootstrap sampling (WBBS) is used to select intervals, and the mean of the absolute values of regression coefficients of the corresponding interval determines the weights of sub-intervals. Third, significant multivariate competitive population analysis (SMCPA) that combines the ideas of substantial multivariate correlation (SMC) and MPA, and employs WBS is an improved version of bootstrap sampling with different weights on sampling objects and exponential decline function (EDF) competition method used to force the elimination of uninformative or redundancy variables [7]. For corn and wheat protein datasets, both methods select informative intervals including the BOSS. However, the BOSS was unstable, and only a few variables are selected compared with other high-performance methods that were more accurate and selected more variables in these crucial intervals.

In terms of the selection of spectra intervals, all models except FOSS (i.e. BOSS, SBOSS, and SMCPA) have not considered this method, although it can provide a reasonable interpretation. Thus, using this method in the proposed model is expected to improve the accuracy as the vibrational spectral band relating to the chemical group generally has a width of 4–200 cm^{-1} [6]. Besides, none of these approaches, including FOSS, searches for optimal combinations in specific informative intervals.

Therefore, in this study, a new hybrid model is proposed based on the BOSS method. However, the proposed hybrid model works by incrementing the number of variables being selected rather than decreasing them. To the best of the authors' knowledge, there is no such hybrid model in the literature based on increasing the number of variables, but there are many developed hybrid models based on reducing the number of variables such as a hybrid VCPA-IRIV model [9], competitive adaptive reweighted sampling-successive projections algorithm (CARS-SPA) [10], and a combination strategy of random forest and backpropagation network (RF-BPN) [11]. The mentioned methods have their own merits and unique characteristics. The decreased-number-based variable selection methods attempt to utilize the features of other methods by making an effective combination. However, the overall performance can be reduced significantly if the preliminary method does not successfully select the key variables [12]. The proposed hybrid method follows the same concept by taking the advantage of the BOSS method that successfully proved to select important intervals: however the BOSS method selects fewer variables and does not select optimal combinations, so we used IRVS to add more variables in these intervals to have an excellent performance. Besides, we focus on the importance of interval as proved to be more robust and more interpretable, so we develop our model that increases the numbers in those informative intervals. The disadvantage of the increased number of variable selection methods that we don't know what is the optimal number of variables that need to be increased. Therefore, there is a need to tune the parameter to decide the optimal number.

The novelty in this research is the following:

1-There is no previous hybrid variable selection method in NIR spectroscopy based on increasing the number of variables. However, most of the studies use a hybrid model to eliminate variables. This paper introduces a new hybrid method based on the incremental approach.

2- The number of datasets used in the evaluation of the proposed hybrid model (i.e. 7 NIR datasets) is considerably large, which led to a proper evaluation. The used NIR datasets are corn datasets with moisture, oil and protein properties, hemoglobin, diesel fuel with total aromatics properties, soy with moisture properties, and wheat protein datasets.

3- Investigating the proposed hybrid methods with two high-performance model include hybrid VCPA-IRIV and BOSS.

4- Providing a comprehensive review of different variable selection methods in terms of the ability to select informative intervals and the performance of the models and numbers of the chosen variable.

The remainder of this paper is divided into the following sections. Related studies are described in Section II, followed by a detailed description of the proposed hybrid method in Section III. The datasets used in this study are described in Section IV. The experimental work and obtained results are presented in Section V. Finally, the conclusion of this study is presented in Section VI.

II. RELATED WORKS

During the last several decades, a large number of various mathematical strategies for variable selection have been employed in NIR spectroscopy.

Li-Li Wang has classified the single variable selection methods and interval variable selection methods into a different classification [13]. The only variable selection methods have been classified into classic stepwise methods, variable raking-based strategy, penalty-based strategy, MPA, heuristic algorithm-based strategy, and some other methods include successive projection algorithm (SPA) and uninformative variable elimination (UVE). On the other hand, the interval selection method is classified into; (1) classic methods including interval PLS (iPLS) and its variants, (2) moving windows PLS (MWPLS), and its variants; (3) penalty-based methods include elastic net combined with partial least squares regression (EN-PLSR), iterative rank PLS regression coefficient screening (EN-IRRCS) and group PLS (gPLS); (4) sampling-based methods include iPLS-Bootstrap and Bootstrap variable importance in projection (Bootstrap-VIP); (5) correlation-based method include sure independence screening and interval PLS (SIS-iPLS); finally, (6) projection-based methods include interval successive projections algorithm (iSPA).

The MPA method has been widely used as it shows a promising prediction ability. The MPA has been classified into single variable model population analysis and interval model population analysis. The former includes random frog (RF) [14], iteratively retains informative variables (IRIV) [15], variable iterative space shrinkage approach (VISSA) [16], iteratively variable subset optimization (IVSO) [17], CARS [18], stability competitive adaptive reweighted sampling (SCARS) [19], sampling error profile analysis LASSO (SEPA-LASSO) [20], BOSS [4] and SBOSS [5]; while the latter includes interval random frog (iRF) [21], interval variable iterative space shrinkage approach (iVISSA) [22], interval combination optimization (ICO) [23] and fisher optimal subspace shrinkage (FOSS) [6].

Moreover, selecting the variables on near-infrared spectroscopy by utilizing models that hybridize two or more different techniques was recommended in [12]. In particular, the UVE method was used in [24] to filters the noise variables: then the SPA method was used to achieve an excellent selection. It is known as the UVE-SPA-MLR hybrid model. Another hybrid model called iPLS-mIPW combined two methods, i.e., iPLS with mIPW [25]. In iPLS-mIPW, the informative intervals were obtained using the iPLS method initially. Then further variables selection was performed using mIPW. Additionally, to select critical wavelengths in NIR spectra, the random forest was hybridized with the BP network by Chen et al. [11]. In the proposed model, some informative wavelengths initially selected using random forest. Then a new comprehensive variable group is produced, using BP network, with minimum errors. Recently, a VCPA-based hybrid model was proposed by Yun et al. [9]. In this model, VCPA was hybridized with the genetic algorithm (GA) and IRIV separately. Firstly, VCPA was used to continuously shrink and optimize the variable space from big to small. After that, additional optimization was performed, on the variables remained by VCPA, using IRIV and GA.

Table 1 shows the comparison between previous methods in terms of selecting informative intervals and the performance of the methods and the number of the variable selected.

III.PROPOSED MODEL

In this section, a description of the proposed hybrid method named bootstrapping soft shrinkage approach and interval random variable selection (BOSS-IRVS) is provided in detail. It combines both the choice of informative intervals using the BOSS method, as illustrated in Section A, and an interval variable selection method, as shown in Section B. Then, a brief description of the compared methods and the model validation is given in Section C and D, respectively. Besides, Figure 2 shows an illustration of the proposed model.

A. Informative intervals selection using BOSS methods

The BOSS approach is designed to choose informative intervals, and that happens with the existence of collinearity. In a suitable shrinkage manner, data from regression coefficients are used by this approach [26]–[29]. Two types of sampling methods are used, including Bootstrap sampling (BSS) and Weighted Bootstrap (WBS). The purpose of the sampling method is to produce a random combination of variables and to construct sub-models of the system. Thus, two methods are coupled and used, including MPA [25] and PLS regression [29], to extract the information from the sub-models. The BOSS method has five main steps to select the informative intervals illustrated as follows.

Step 1: BSS is used to produce *K* subsets on a variable space. The variables chosen for BSS are extracted from each dataset, and the redundant variables are excluded. Thus, only the unique variables have remained. The replacement number, in BSS, is identical to the total number of variables *P*. Therefore, the number of variables chosen is roughly $0.632\,P$ in each subset. Here, all variables must be treated equally so that they can be picked into subsets with the same probability, i.e., equal weights (w) are set for all variables.

Step 2: The subsets obtained are used to construct *K* PLS sub-models. Then, the prediction error is calculated based on RMSEV, and a percentage of the lowest RMSEV models is selected, representing the best models (e.g., 10 percent).

Step 3: Regression coefficients (RC) are computed and adjusted to the absolute value of all elements on the regression vector and normalize each regression vector to unit length for any extracted model. Subsequently, equation (1) is used to obtain new weights for variables by summing up the normalized regression vector.

$$w_i = \sum_{A=1}^{K} b_{i,A} \tag{1}$$

where $\mathrm{w_i}$ is the new weight for *i*th variable, *K* denotes the number of sub-models and $\mathrm{b}_{\mathrm{i},A}$ represents the absolute normalized regression coefficient value for the *i*th variable in the *A*th sub-model.

Step 4: The WBS generates new subsets using WBS according to the variables' new weights. As in BSS, the variables chosen are extracted in each dataset to construct the sub-models, and the redundant variables are excluded. The average number of variables calculated in Step 3 is used to determine the number of replacements in WBS. Therefore, in the new subsets, the number of variables is 0.632 times of those previously determined [4]. The aim behind this step is to guarantee that the variables with larger absolute values of regression coefficients are likely to be selected in the best sub-models.

Step 5: Repeat Step 2-4 until a number of variables in the new subsets are 1, then return the optimal subset, which has the lowest RMSEV.

Step 6: Repeat the BOSS method twenty times to select informative intervals.

B. Selection of informative variables in informative intervals

After applying the BOSS method to NIR datasets to select informative intervals by Algorithm 1, the output of this algorithm will act as the input for Algorithm 2. The later algorithm will select informative variables in the informative intervals. The selection of informative variables is affected by three parameters that need to be tuned carefully. These parameters are:

(i) The number of populations (*np*):

To select an adequate number of populations, three cases of 50, 100, 500 populations were investigated. For example, 50 populations combine 50 individuals in which each individual combines the variables selected in Algorithm 1 and the interval random variables method, which search for informative variables in informative intervals. Five hundred populations were chosen as the optimized number of populations based on 20 replicated results shown in Figure 3. Therefore, 500 population was set in this work. From Figure 3 (a), it should be noted that when the 50 generations have been used, the value of RMSEC varies from 3.1 to 3.9, which is an indication of underfitting, as shown in Figure 3. (b). However, with 500 generations, both values of RMSEC and RMSEP are dropped to the lowest level, which avoids overfitting and gives the best performance compared with 50 and 100 populations.

(ii) The way of selecting random variables:

The first choice is to choose random variables gradually or to select random variables at one time. Selection of random variables gradually means to select specific random variables in each run while selecting random variables at one time means to select all the random variables in only one run. Every random variable has small random interval from the big interval selected by BOSS. Figure 4 proves that the gradual selection of random variables is the optimal approach, which avoids overfitting. From the same figure, it can be seen that selecting random variables at one time leads to low RMSEC and high RMSEP, while gradual selection leads to low RMSEP.

(iii) The number of informative variables selected (*nv*):

To select an adequate number of added informative variables, three cases of 3, 6, 9 variables were investigated. Among the three numbers tested shown in Figure 5, it can be realized that the three variables have the worst RMSEC value, while the nine variables have the best values. However, with nine variables being selected, the RMSEP is high. As a compromise, the 6 number of variables is chosen as it produces the lowest RMSEP value and an acceptable RMSEC value.

The pseudocode of the proposed algorithm is presented in **Algorithm 2**. In the beginning, five hundred random populations are generated. Each individual in the population combines input variables and three random variables from the informative intervals. The input is the variables selected in **Algorithm 1**. For each individual, RMSEC value is calculated, and the individual with the lowest RMSEC value is selected. These steps are repeated until *ny* random variables have been selected. For each round, the input is updated by adding the three variables chosen from the previous round.

TABLE 1. **Comparison between previous methods in terms of selecting informative intervals and the performance of the methods and numbers of the variable selected**

Response	Methods	References	The average no of selected variables	Remarks	Performance
Corn moisture	MCUVE	[4]	12.3 ± 1.7	Select a larger number of variables, select informative intervals, and selects variables in the uninformative region	BOSS and GA-PLS outperformed PLS, MCUVE and CARS
	CARS		5.1 ± 2.5	Select informative intervals, select a lower number of variables, lack of concentration around informative intervals compared than BOSS	
	GA-PLS		7.1 ± 3.2	Select informative intervals	
	LASSO	[30]	37	Select informative intervals and select uninformative variables select a larger number of variables	SPPA-LASSO outperformed PLS, PCR, LASSO, MWPLS, OHPL, MC-UVE and SCARS
	MWPLS		119	Fail to select informative intervals, select uninformative variables and select a larger number of variables	
	OHPL		57	Select informative intervals and the select a larger number of variables	
	SCARS		5	Select the informative intervals and select a lower number of variables	
	SEPA-LASSO		7	Select the informative intervals and select a lower number of variables	
	IVSO	[17]	2.3±0.8	Select the informative intervals and select a lower number of variables	IVSO outperformed PLS, CARS, and MC-UVE
	IRIV	[31]	7.7 ± 3.6	Select informative intervals	GA-PLS outperformed IRIV. IRIV outperformed CARS and MC-UVE
	siPLS	[22]	40	Select informative intervals and the interval widths were not optimized Select a larger number of variables	iVISSA outperformed PLS, siPLS, MW-PLS, CARS, GA-PLS and iRF.
	iRF		35.1± 5.5	Select informative intervals, select uninformative intervals and select a larger number of variables	
	iVISSA		14.2± 3.7	Select informative intervals select a larger number of variables	
Corn oil	MCUVE	[4]	87.4 ± 41.6	Select a larger number of variables, select informative intervals selects some uninformative variables	BOSS outperformed PLS, MCUVE, CARS, and GA-PLS
	CARS		16.0 ± 4.8	Select informative intervals, lower number of variables and select uninformative intervals	
	GA-PLS		79.2 ± 31.2	Select a larger number of variables, select informative intervals, and select uninformative intervals and select a larger number of variables	
	GA-iPLS	[9]	59.4 ± 6.8	Select informative intervals and select uninformative intervals	VCPA-GA outperformed VCPA-IRIV, CARS, GA-iPLS and VIP-GA
	VIP-GA		25.2 ± 7.8	Select informative intervals and select uninformative intervals	
	VCPA-GA		32.9 ± 11.4	Select the informative intervals and has a good different variable combination in the informative intervals compared than GA-iPLS	
	FOSS	[6]	45.3± 10.2	Select an informative region and has a good concentration	FOSS outperformed PLS, MW-PLS, iRF, iVISSA and BOSS
Corn protein	MCUVE	[4]	112.2 ± 16.8	Select a larger number of variables, select informative variables, and select intervals between 1900 and 2000 which lower the performance	BOSS outperformed PLS, MCUVE, CARS, and GA-PLS
	CARS		20.2±8.5	Select informative intervals and select intervals around 2300 which lower the performance.	

	GA-PLS		51.6 ± 20.7	Select a larger number of variables and select informative intervals and select intervals around 2300 which lower the performance	
	VISSA	[23]	172	Select variables across all spectra and select a larger number of variables	ICO outperformed PLS, VISSA, iVISSA, VISSA-iPLS and GA-iPLS
	iVISSA		241	Select variables across all spectra and select a larger number of variables	
	VISSA-iPLS		105	Select informative intervals, select some more variables around 1670-1710nm and 2192-2224nm which lower the performance. and select larger number of variables	
	GA-iPLS		70	Select informative intervals	
	ICO		69	Select informative intervals	
	SBOSS	[5]	25±7	Select informative intervals and has a good concentration than BOSS	SBOSS outperformed SCARS, BOSS, CARS, GA-PLS, and MCUVE
Soy moisture	MCUVE	[4]	32.8±21.4	Select two informative intervals, Select other variables in other intervals and select larger number of variables	BOSS outperformed PLS, MCUVE, CARS, and GA-PLS
	CARS		6.4±4.6	Select two informative intervals Select other variables in other intervals	
	GA-PLS		18.3±6.0	Select two informative intervals, and Select other variables in other intervals	
	siPLS	[22]	21	Select two informative intervals, and select the wavelengths around 1520 nm	iVISSA outperformed PLS, siPLS, MW-PLS, CARS, GA-PLS and iRF
	iRF		29.4±6.4	Select two informative intervals and select around 2480 to 2500 nm	
	iVISSA		25.5±0.9	Select two informative intervals and select around 2480 to 2500 nm	
	MW-PLS		48	Select two informative and select around 2400 nm and select larger number of variables	
Total Diesel Fuel	MCUVE	[4]	107.4±54	Select variables between 1450 and 1550 and between 1200 and 1300 and between 800 and 1200	BOSS outperformed MCUVE, CARS, and GA-PLS
	CARS		28.8±10.3	Select variables between 1450 and 1550 and between 1200 and 1300 and between 950 and 110	
	GA-PLS		87.9±44	Select variables between 1450 and 1550 and between 1200 and 1300 and between 950 and 110	
Wheat protein	MC-UVE	[17]	10.6 ± 1.3	Select informative intervals and select many variables in other intervals	IVSO outperformed PLS, CARS, and MC-UVE
	CARS		9.8 ± 2.8	Select informative intervals Selects many variables in uninformative intervals	
	IVSO		14.8 ± 3.0	Select informative variables around 1144-1296nm	
	GA-PLS-LRC	[32]	19±5	Select informative variables around region 1100–1340 nm	GA-PLS-LRC outperformed GA-PLS
	VCPA	[7]	8.9±1	Select informative variables around 1150–1350 nm	SMCPA outperformed VCPA, CARS, and BOSS
	SMCPA		6.7±0.7	Select informative variables around 1150–1350 nm and has good selection of variables in informative intervals	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3023681, IEEE Access

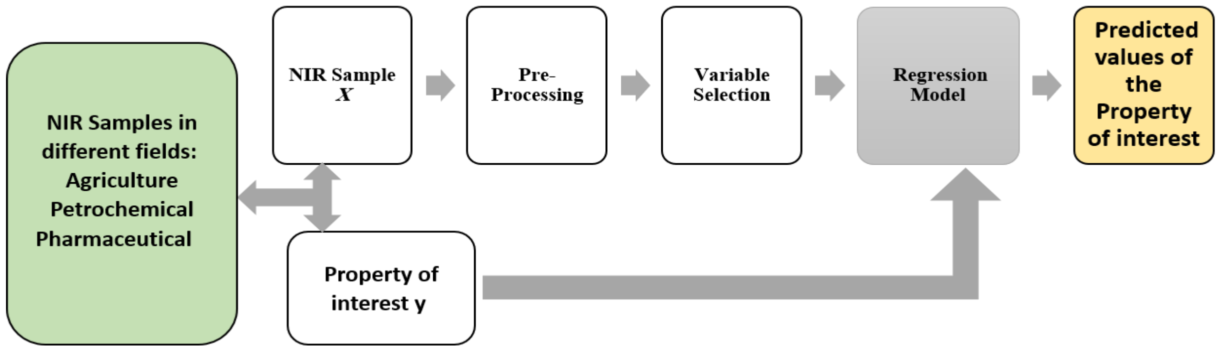

NIR Samples in different fields: Agriculture Petrochemical Pharmaceutical
NIR Sample X
Pre-Processing
Variable Selection
Regression Model
Predicted values of the Property of interest
Property of interest y

Figure 1. **The process of multivariate analysis for near infrared spectroscopy**

Algorithm1: Selection of informative intervals using BOSS

```
Input data: **X** [N, P], y [N+1]
Set the maximum number of iterations (NI), bootstrap resample size (N), number of variables (P), and number of subsets (K).
Set a sampling method for BSS.
Generate K subsets using BSS: all the variables are with equal weights (w).
Assign equal weights (w) for the generated variables.
Set RMSEV to zeros.
retained_variables=P.
j=1
While (j<=100 OR retained_sebsets > 1)
begin
    Build KPLS sub-models using the subsets obtained.
    Calculate RMSEV of the sub-models.
    Extract best models with the lowest RMSEV.
    Calculate regression coefficients for each extracted model.
    Change all the elements in the regression vectorRV to absolute value.
    Normalize each RV to have unit length.
    Sum up the normalized RV to obtain new Ws for variables using equation (1).
    Apply WBS according to the new Ws for variables to generate new subsets.
    Extract the unique variables to build up the sub-models.
    Calculate the average (avg) of the extracted variables.
    Determine the number of replacements in WBS using avg.
    Compute and retain the variables with the most considerableabsolute value of regression coefficients
end while
Apply 5-fold cross-validation
to analyze the N variable subsets statistically.
Choose the Variable subset with minimum RMSEV as the optimal variable subset.
Run the BOSS twenty times to select informativeintervals.
```

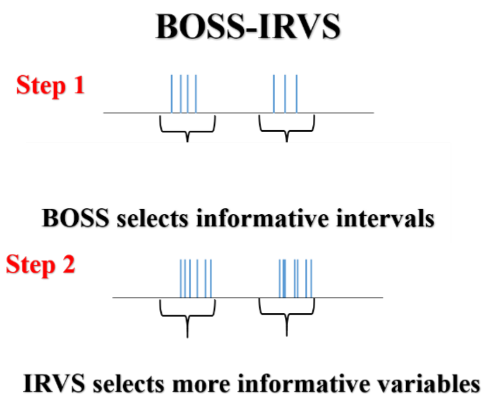

BOSS-IRVS
Step 1
BOSS selects informative intervals
Step 2
IRVS selects more informative variables

Figure 2. **illustration of the BOSS-IRVS model**

Algorithm2: Search for informative variables in informative intervals

```
Input: Variable selected by BOSS
Output: Select informative variables in informative intervals
$r = 0$ // r is the # of random variables selected
While ($r \neq nv$)
Begin
    Determine the intervals selected by BOSS in **Algorithm1**
    Generate 500 random population (np)
    Individual = Input + three random variables from intervals
    selected by BOSS
    Calculate RMSEC of each individual
    Select the individual that has the lowest RMSEC
    Input = the individual that has lowest RMSEC
    $r = r + 3$
End while
```

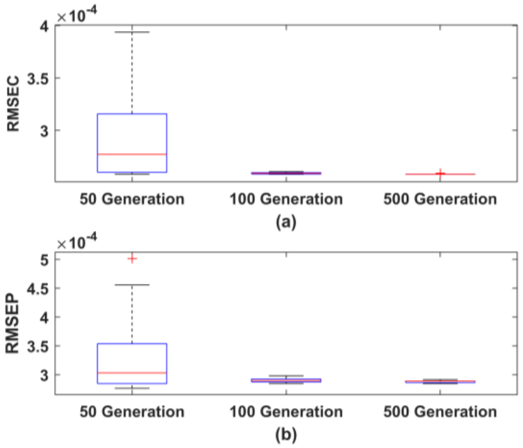
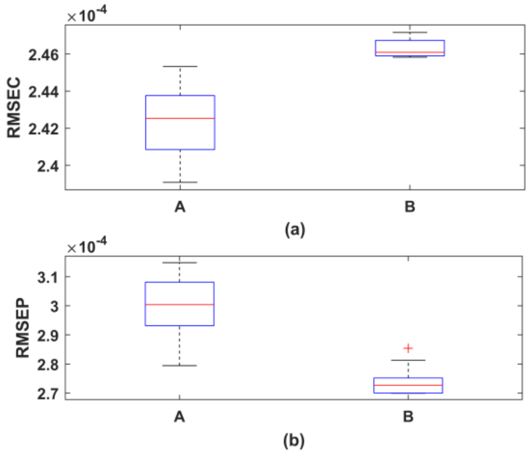


Figure 3. **Boxplot of 20 times for investigating the effect of number of populations on moisture corn dataset (a) RMSEC (b) RMSEP**



A- One-time selection method
B- Gradual selection method

Figure 4. **Boxplot of 20 times for investigating the effect of selecting six random variables by one-time and gradual selection method on moisture corn dataset, where (a) is RMSEC and (b) is RMSEP results.**

C. An outline of the Hybrid VCPA-IRIV compared methods

The VCPA-based hybrid variable selection technique was recently proposed by Yun *et al*. The concept of continuous shrinkage of variable space is the fundamental idea of the original VCPA method. The proposed hybrid VCPA method has two main phases. In the first phase, a modified VCPA was used to shrink the variable space continuously from big to small and optimizes it. For further optimization, the IRIV method was applied in the second phase.

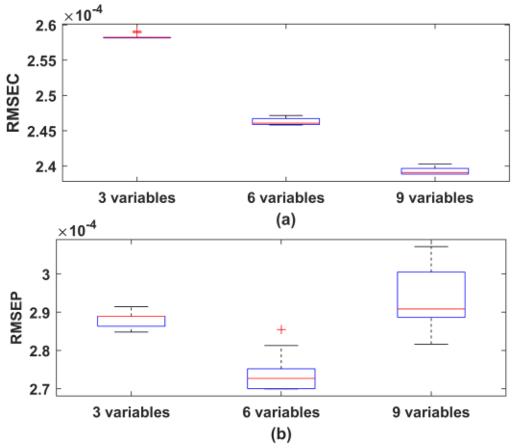


Figure 5. **Boxplot of 20 times for investigating the effect of selected 3 variables,6 variables, 9 variables on moisture corn dataset (a) RMSEC (b) RMSEP**

D. MODEL VALIDATION

With 5-fold cross-validation and test sets, the predictive ability of the models is assessed by the root mean squared error of training (RMSEC), the root mean squared error of cross-validation (RMSEV), the root mean squared error of prediction (RMSEP), the coefficient of determination of training (Q^2_C), the coefficient of determination of cross-validation (Q^2cv) and the coefficient of determination of test set (Q^2_T).

$$RMSEC = \sqrt{\frac{\sum_{I=1}^{Ntrain}\left(y_i - \hat{y}\right)^2}{Ntrain}} \tag{2}$$

$$Q_C^2 = \frac{\sum_{I=1}^{Ntrain}(y_i - \hat{y})^2}{\sum_{I=1}^{Ntrain}(y_i - \bar{y}_i)^2} \tag{3}$$

where y_i ,$\hat{y}$, and $\bar{y}_i$ are the experimental, predicted, and the average of predicted properties, respectively. $Ntrain$ is the number of calibration samples in the training set. The RMSEP and RMSEV are computed similarly as RMSEC, while $Q^2_T$and Q^2CV are computed as and Q^2_C, but with different $Ntrain$ values that are changed with the testing sample for RMSEP and Q^2_T only.

IV. DATASETS

In this study, seven NIR datasets have been used to evaluate the BOSS-IRVS, which are datasets of diesel, soy, wheat protein, corn, and hemoglobin. The important details of these datasets are summarized below.

A. Corn datasets

From *http:/www.eigenvector.com/data/Corn/index.html*, four NIR corn datasets were collected. In each dataset, there are 80 corn samples measured by m5 NIR spectrometers. Also, there are 700 wavelength points of 2 nm intervals in the range of 1100-2498 nm for each spectrum. The properties of interest were used are oil, protein, and the content of moisture. The samples were divided, in each of the 80 corn samples, equally into a 60 training set and a 20 independent test set.

B. Diesel fuels dataset

This dataset has been downloaded from the website *http:/www.eigenvector.com/data/SWRI/index.html*. The range of wavelength points is between 750-1550 nm at intervals of 2 nm for each spectrum, including 401 points. Only one property of interest is considered, which is the total aromatics, while the remaining properties are removed. The 20 high-leverage samples and one of the two random samples were used for each dataset to create the training set. The other group was used as an independent test set, leading in total dataset sample partitions 138 and 118 for training and testing, respectively.

C. Soy datasets

Spectrometer NIR was used to measure the samples of soy flour [33]. There are 175 wavelengths in each spectrum, with 8 nm in the range of 1104 and 2496 nm. The moisture content was considered as properties of interest. According to the reference [33], each dataset contains 54 samples, split between the training set (40 samples) and the test set (14 samples).

D. Wheat dataset

This NIR dataset [34] contains 100 wheat samples. The spectrum was reported at intervals of 2 nm from 1100 to 2500 nm with a spectrum of 701 points. The property of interest *y* is the protein value. Owning the problem of ‘large p, small n’ [35][36], an acceptable window size compresses the original spectrum into a limit of 200 frames [37]. This dataset is reduced to 175 variables by limiting window size to 4, and each of the original four variables is averaged. Out of 100 samples, 80 was used for training and 20 for testing.

E. Hemoglobin dataset

Using the IDRC shootout 2010 software, Karl Norris [38] has produced this dataset that has been used by Mohd Nazrul Idrus [39]. With the spectrometer of NIR Systems 6500, the blood samples have been analyzed. The blood hemoglobin reference was measured by a high-volume hematology analyzer. All spectra have 700 variables of 2 nm interval in the range between 1100 and 2498 nm. To evaluate the model, the dataset is divided into 173 sets and 194 unseen data sets, respectively, for training, and blind testing to measure the model’s predictive accuracy.

V. Results and Discussions

To assess the performance of the BOSS-IRVS, some high-performance wavelength selection methods, including BOSS and VCPA-IRIV, are used for comparison. All codes were applied in Matlab. The datasets are centered. In this study, the calibration set is used for building the model and performing the variable selection. The independent test set is then used to validate the calibration model. Several evaluation metrics, such as the RMSEV, Q^2_cv, RMSEC, Q^2_C, RMSEP, and Q^2_T, are used to measure the performance of the introduced model. At the same time, the maximum number of latent variables (mnLV) and the number of selected variables (nVAR) are also calculated. Each method is repeated 20 times to ensure the reproducibility and stability of the evaluation. The parameter setting for VCPA-IRIV are as follows: α = 20 which is the mean number of each BMS sampling, *EDF_run* = 50 which is the number of exponentially decreasing function (EDF) run, *BMS_run* = 1000 which is the number of BMS run, σ = 0.1 which is the ratio of the best minus worst models of *K*sub-models, *L* = 100 which is the number of the left variables in the final run of EDF, *A_max* = 10 which is the maximal principle component to extract for PLS, *fold* = 5 which is the group number of cross-validation, and *method* = center which is the pretreatment method. In respect to the last three setting parameters, BOSS has similar settings as VCPA-IRIV. Last, the number of bootstrap used in BOSS, num_bootstrap is set to 1000.

A. Corn dataset

The results of variable selection methods, i.e. VCPA-IRIV, BOSS and BOSS-IRVS, on moisture, oil, and protein properties of corn datasets are summarized in Table 2. The results show that, on the three datasets, the BOSS-IRVS outperformed the prediction ability of the BOSS and the hybrid model of VCPA-IRIV. In detail, using BOSS-IRVS, the values of RMSEP for moisture datasets are improved from 3.2328e-04 to 2.8804e-04 when three variables are added and to 2.7360e-4 when six variables are added. For the oil dataset, the values of RMSEP are improved from 0.0347 to 0.0197, and 0.0174 with three variables and six variables are added, respectively. For the protein dataset, the values of RMSEP are improved from 0.0322 to 0.0192 and 0.0159 when three variables and six variables are added, respectively. In terms of the VCPA-IRIV model. The RMSEP values are 3.2341e-04, 0.0201, and 0.0272 for moisture, oil, and protein respectively; while for BOSS-IRVS model, they are 2.7360e-04, 0.0174, and 0.0159.

The variables selected by different selection methods on moisture datasets are shown in Figure 6. The wavelengths chosen by BOSS, VCPA-IRIV, and BOSS-IRVS models are located in two intervals and selected the two wavelengths of 1908 nm and 2108 nm. These two wavelengths are regarded as the key wavelength by Li et al. [20], [9], which correspond to the water absorption and the combination of O-H bonds according to the literature [22]. The number of the variable selected by the BOSS is 3.8, which indicates that the BOSS algorithm misses important variables and ignores the high correlation among consecutive variables. The BOSS-IRVS improved the BOSS prediction ability by adding six important variables. The VCPA-IRIV selects 5.5 variables which are the same as the BOSS-IRIV model when three variables are added. However, the variables selected by the BOSS-IRVS model give better performance compared to the variables selected by VCPA-IRIV and the reason is that the variable combinations of the BOSS-IRVS are better than the variable combinations of VCPA-IRIV. For oil dataset, From the Figure 7, it can be observed that VCPA-IRIV, BOSS and BOSS-IRVS methods select informative spectra intervals near 1700 nm (region 1) and 2300 nm (region 2), which correspond to the second and first overtones of the C-H stretching mode and the combination of C-H vibrations [8]. The VCP-IRIV shows a good concentration on the two intervals compared with the BOSS method, which the variables selected by BOSS is unstable since it uses bootstrap sampling. The BOSS-IRVS combines the variables selected by the BOSS and added six variables in the informative intervals selected by BOSS, which lead to outperforming VCPA-IRIV model. The BOSS has the lowest variables selected then both VCPA-IRIV and BOSS-IRVS models have the same number of the variable selected. For the protein dataset, From Figure 8, we could observe that VCPA-IRIV, BOSS, and the BOSS-IRVS methods select the combination of several groups that are chemical meaningful for data analysis of spectrum [5]. All the methods selected the intervals around 1680, 1800 and 2180 nm. It can be noticed that these selected intervals cover a wide range linking to the complicated structure of the protein, e.g. C-H, O-H and N-H bond with different vibration pattern, complex microenvironment of the three bonds, and the interaction of them [4]. The lowest number is selected by BOSS followed by the BOSS-IRVS model with three added variables, and then both the VCPA-IRIV and the BOSS-IRVS with six added variables have nearly the same variables selected. The BOSS-IRVS selects important variables near to intervals 1800 and 2180, which outperformed the BOSS and VCPA-IRIV.

Furthermore, from Figure 6, it can be seen that the proposed model had high stability variables since it focuses on specific important intervals. In more detail, all variables selected by BOSS in the first step of the proposed model are considered informative variables. Then, the selected BOSS variables are used as input for the IRVS algorithm, which means that the IRVS algorithm chooses the same variables selected by BOSS and adds the six selected incremental variables to them. The process is repeated 20 times until the optimal incremental number of variables is reached. As a result, the variables selected in the first step will always have the highest frequency of 20.

From Table1, with respect to moisture dataset, there are many methods that succeed to select informative intervals include CARS, MCUVE, OHPL, SCARS, SPEA-LASSO, BOSS and VCPA-IRIV. However, some methods have lower performance compared to other methods due to various reasons. For instance, select uninformative variables in other intervals such as in iRF and LASSO methods, or low concentrate when choosing variables in informative intervals such as in CARS. Furthermore, although the methods succeeded to select important intervals, it chooses many variables, including uninformative variables such as in OHPL. Moreover, the combinations of variables are different, which the reason why some methods outperformed other methods that select the same informative intervals, such as SPEA-LASSO outperformed SCARS. Our hybrid method succeeded to select a lower number and good concentration in the informative intervals by select informative variables in informative intervals. For the oil dataset, some methods succeed to select informative intervals but select uninformative variables such as CARS, GA-PLS and VIP-GA. GA-pills and VCPA-IRIV succeed to select

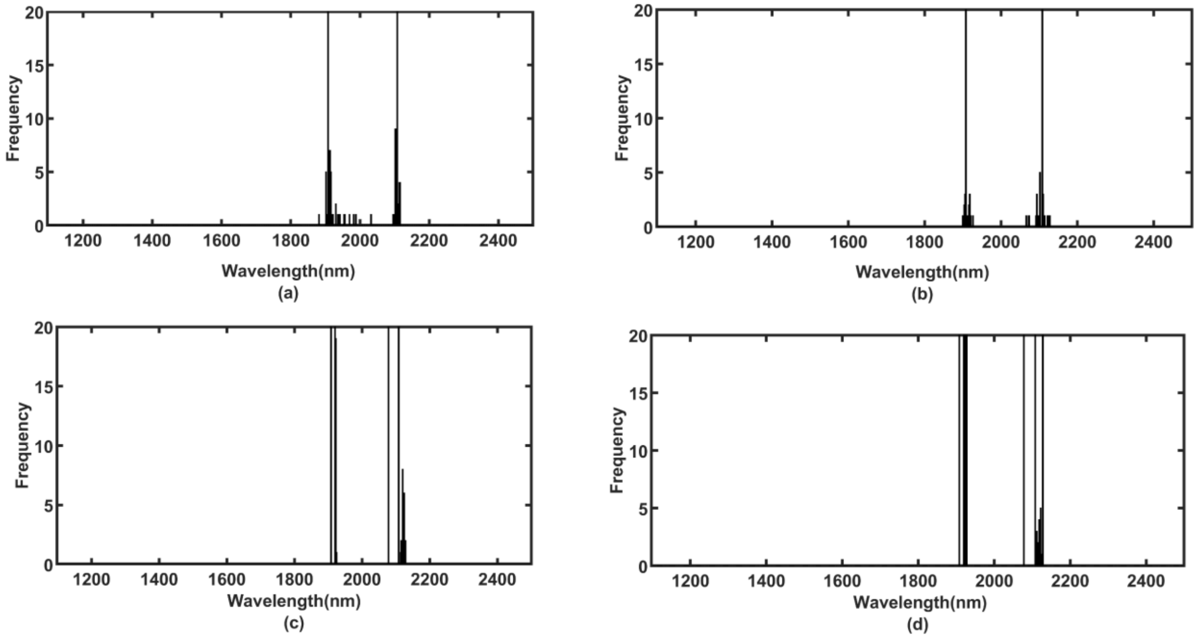


FIGURE 6. **The frequency of selected variables within 20 times on the Corn Moisture dataset: (A) VCPA-IRIV, (B) BOSS, (C) proposed method selected three variables (D) proposed methodsselected six variables**

informative variables; however VCPA-IRIV chooses the optimal number of variables and has a good combination of variables which outperformed GA-iPLS. FOSS method has a good performance by succeeding to concentrate in informative intervals and choose informative variables in theses informative intervals. Our hybrid method succeeded to select a lower number of variables and select informative variables in informative intervals. For the protein dataset, CARS and GA-PLS select important intervals; however, they also select uninformative intervals. The selection of uninformative intervals reduces the performance of both CARS and GA-PLS. VISSA and iVISSA methods have low performance because they select variables around all spectra. ICO method outperformed CARS, MC-UVE, VISSA and iVISSA because of the low number of variables and succeeded to select variables in informative intervals. A recent paper called SBOSS outperformed SCARS, BOSS, CARS, GA-PLS, and MCUVE. The SBOSS has a low variable with a good selection of variables in the informative interval.

TABLE 2. **Results for the Corn datasets. nVAR: number of variables; mnLVs: max number of latent variables; RMSEC: root mean-square error of calibration RMSEV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; coefficient of determination of calibration; Q2_C; Q2_CV: coefficient of determination of cross-validation; Q2_Tcoefficient of determination of test set**

Response	Metrics	VCPA-IRIV	BOSS	BOSS-IRVS with three variables added	BOSS-IRVS with six variables added
Moisture	mnLV	8	10	6	9
	nVAR	5.5±2.1	3.8±3.1	6	9
	RMSEC	2.6775e-04±0.0000	2.7252e-04±0.0000	2.5836e-04±0.0000	2.4625e-04±0.0000
	RMSEV	2.9333e-04±0.0000	2.9253e-04±0.0000	-	-
	RMSEP	3.2341e-04±0.0000	3.2328e-04±0.0000	2.8804e-04±0.0000	2.7360e-04±0.0000
	Q2_C	1.0000±0.0000	1.0000±0.0000	1.000±0.0000	1.000±0.0000
	Q2_CV	1.0000±0.0000	1.0000±0.0000	-	-
	Q2_T	1.0000±0.0000	1.0000±0.0000	1.000±0.0000	1.000±0.0000
Oil	mnLV	10	10	10	10
	nVAR	20.7±5.5	13.5±3.8	20	23
	RMSEC	0.0121±0.0017	0.0242±0.0082	0.0135±2.9847e-04	0.0122±1.8141e-04
	RMSEV	0.0151±0.0023	0.0296±0.0095	-	-
	RMSEP	0.0201±0.0031	0.0347±0.0116	0.0197±0.0014	0.0174±0.0012
	Q2_C	0.9957±0.0011	0.9811±0.0132	0.9947±2.3637e-04	0.9957±1.2837e-04
	Q2_CV	0.9932±0.0021	0.9720 0.0187	-	-
	Q2_T	0.9766±0.0065	0.9240±0.0511	0.9777± 0.0032	0.9826±0.0024
Protein	mnLV	10	10	10	10
	nVAR	29.4±5.8	17.4±3.0	26	29
	RMSEC	0.0136±0.0074	0.0200± 0.0018	0.0110±2.4524e-04	0.0096±1.0164e-04
	RMSEV	0.0187±0.0103	0.0251± 0.0017	-	-
	RMSEP	0.0272±0.0150	0.0322± 0.0039	0.0192 ± 0.0014	0.0159±0.0014
	Q2_C	0.9990±9.5875e-04	0.9984±2.9310e-04	0.9995±2.2038e-05	0.9996±7.9670e-06
	Q2_CV	0.9982±0.0019	0.9974±3.5558e-04	-	-
	Q2_T	0.9961±0.0043	0.9957±0.0010	0.9985±2.2209e-04	0.9990±1.8491e-04

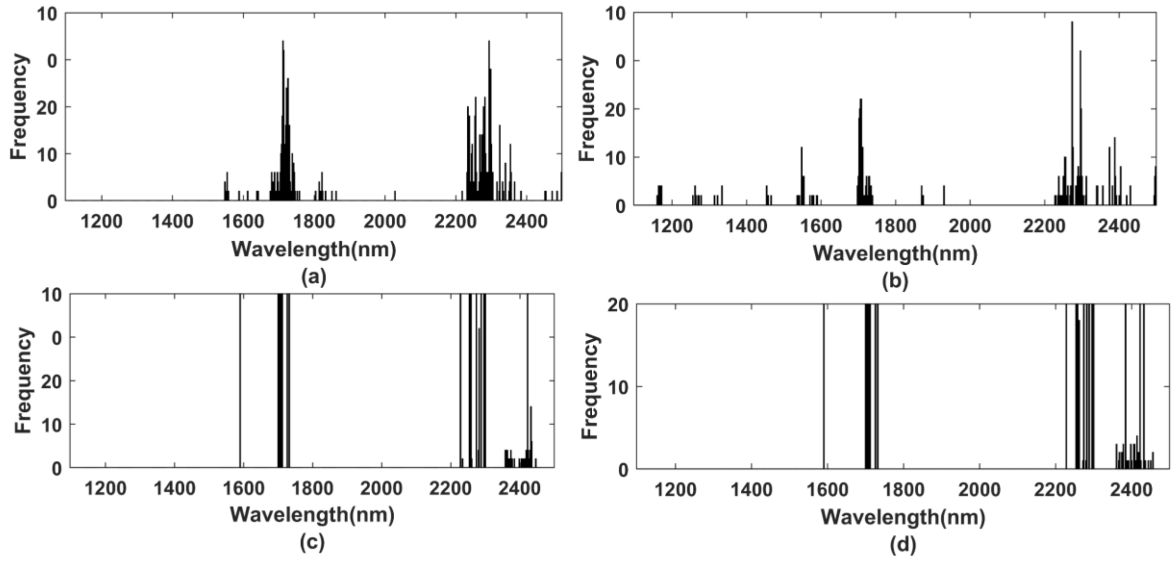


FIGURE 7. **The frequency of selected variables within 20 times on the Corn oil dataset: (A) VCPA-IRIV, (B) BOSS, (C) proposed method selected three variables (d) proposed method selected 6 variables.**

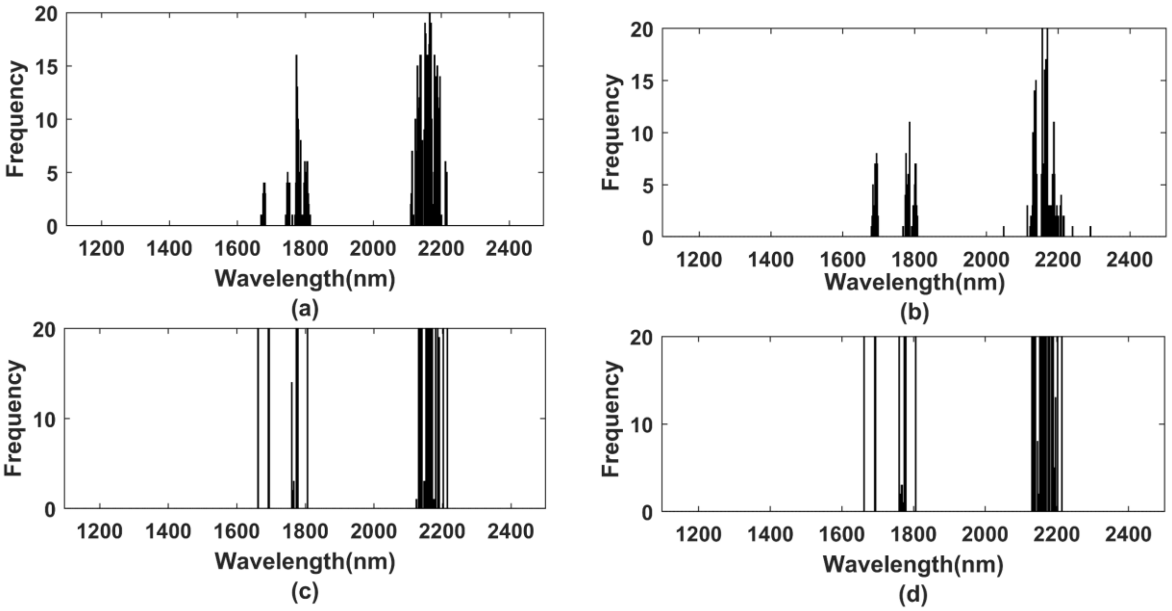


FIGURE 8. **The frequency of selected variables within 20 times on the Corn protein dataset: (A) VCPA-IRIV, (B) BOSS, (C) proposed method selected three variables (D) proposed method selected six variables.**

TABLE 3. **Results for the Soy moisture dataset. nVAR: number of variables; mnLVs: max number of latent variables; RMSEC: root mean-square error of calibration RMSEV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; coefficient of determination of calibration; Q2_C; Q2_CV: coefficient of determination of cross-validation; Q2_T coefficient of determination of test set**

Metrics	VCPA-IRIV	BOSS	BOSS-IRVS with three variables added	BOSS-IRVS with six variables added
mnLV	4	5	5	5
nVAR	4.7±1.6	5±0.7	8	11
RMSEC	0.7067±0.0071	0.6344±0.0196	0.5775±0.0142	0.5804±0.0121
RMSEV	0.7306±0.0049	0.6972±0.0162	-	-
RMSEP	0.9854± 0.0412	0.9126±0.0336	0.8701±0.0408	0.8610±0.0405
Q2_C	0.9341±0.0013	0.9468±0.0033	0.9560±0.0021	0.9555±0.0018
Q2_CV	0.9296±9.5079e-04	0.9358±0.0030	-	-
Q2_T	0.9091±0.0078	0.9126±0.0336	0.9291±0.0066	0.9306±0.0065

B. Soy moisture dataset

The results of variable selection methods on soy datasets are shown in Table 3. A clear ranking of the VCPA-IRIV, BOSS, and the BOSS-IRVS models are as follows. The BOSS-IRVS are followed by BOSS and VCPA-IRIV. The RMSEP for the BOSS-IRVS method with six added

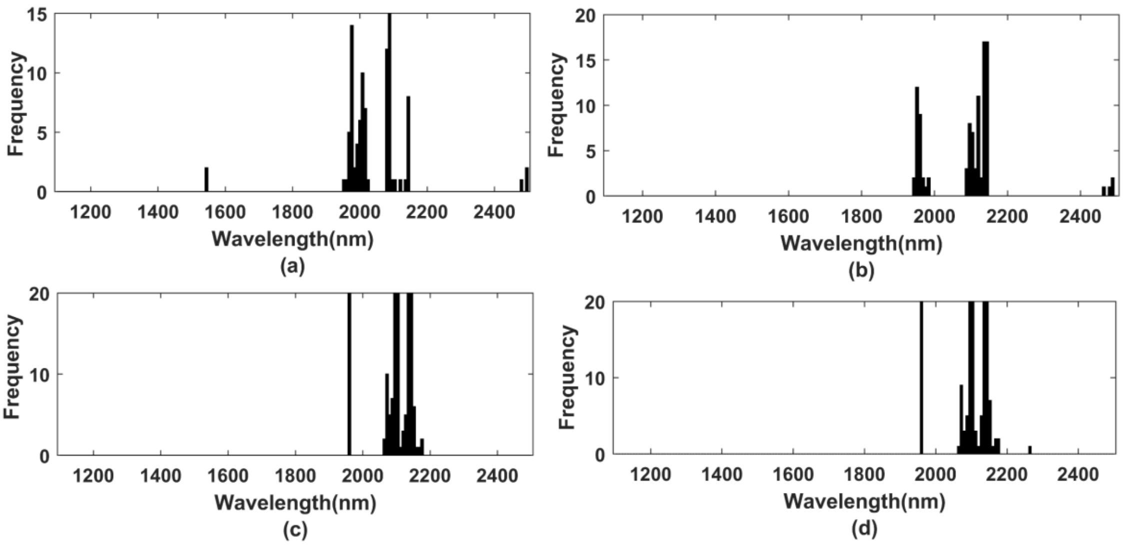


FIGURE 9. **The frequency of selected variables within 20 times on the soy dataset: (A) VCPA-IRIV, (B) BOSS, (C) proposed method selected three variables (D) proposed method selected six variables.**

variables, the BOSS-IRVS with three added variables, BOSS and VCPA-IRIV are 0.8610, 0.8701, 0.9126, and 0.9854, respectively. Moreover, the proposed BOSS-IRVS showed the best Q2_T with 0.9306 compared with 0.9126 and 0.9091 for BOSS and VCPA-RIV, respectively. Figure 9 shows that all the methods select two informative intervals around 1900 nm and 2100 nm, which are selected commonly by four methods. They correspond to the water absorption and the combination of O-H bonds [22]. The VCPA-IRIV selects some variables around 1550 and 2450, and the BOSS method selects intervals around 2450. The BOSS-IRVS method selects variables around 2100 which improves the accuracy of the model. Table 3 and Table1 show the performance of BOSS-IRVS method and other variable selection methods on the soy moisture dataset. Most of these methods select informative intervals; however, some methods select other intervals, such as CARS, MC-UVE, and GA-PLS. Also, some methods select more variables such as MC-UVE, siPLS, MW-PLS, and iRF which show low accuracy compared with a low number of variables such as BOSS and VCPA-IRIV and the BOSS-IRVS methods. The proposed hybrid models select a good combination and an optimal number of variables that achieved higher accuracy.

C. Total diesel fuels dataset

The results of variable selection methods on total diesel fuel datasets are displayed in Table 4 and Figure 10. It shows a clear ranking of prediction ability for all the methods; the BOSS-IRVS with six variables added, the VCPA-IRIV method, BOSS-IRVS with three variables added, and the BOSS method. The values of RMSEP are 0.5965 for the BOSS-IRVS with six added variables, 0.6004 for VCPA-IRV, 0.6026 for BOSS-IRVS with three added variables, and 0.6366 for BOSS. Wavelengths that have been selected by all methods are concentrate in the region of 1000–1100 nm, 1200–1300 nm, and 1450–1550 nm which indicate the importance of these intervals. Moreover, the VCPA-IRIV and BOSS have selected variables around different intervals include intervals between 800 and 900 and between 1300 and 1400. BOSS-IRVS models have selected their variables around these informative intervals which improve the BOSS method significantly. From Table 3 and Table 1, it can be seen that MC-UVE, GA-PLS have a higher number of variables compared to BOSS, CARS, VCPA-IRIV, and proposed hybrid model. The methods that have a low number of variables have a good performance.

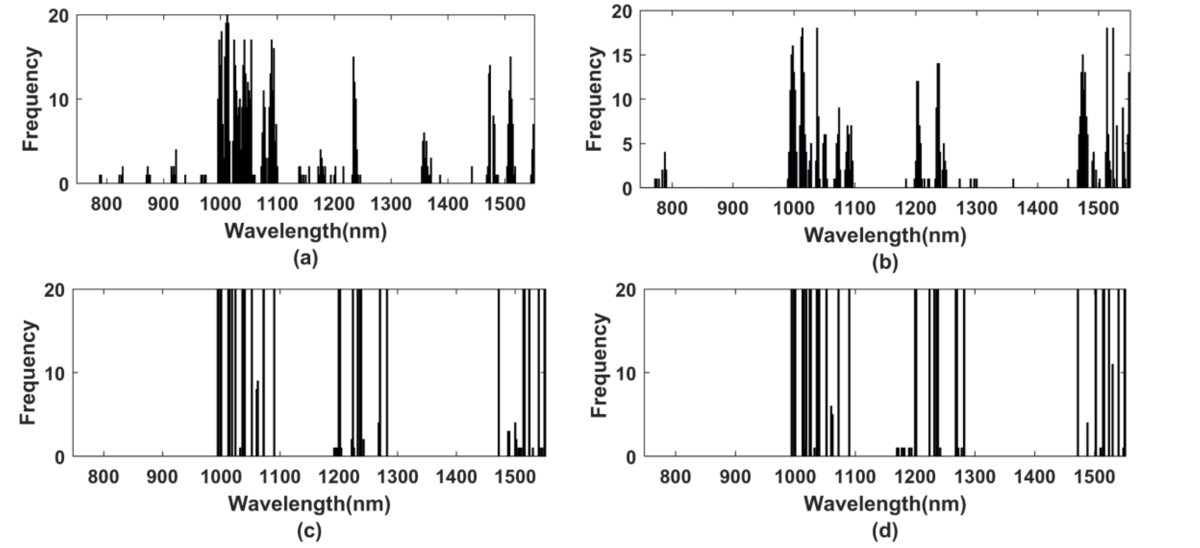


FIGURE 10. **The frequency of selected variables within 20 times on the total diesel fuel properties dataset: (A) VCPA-IRIV, (B) BOSS, (C) proposed method selected three variables (D) proposed method selected six variables.**

TABLE 4 **Results for the total diesel fuel properties dataset. nVAR: number of variables; mnLVs:max number of latent variables; RMSEC: root mean-square error of calibration RMSEV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; coefficient of determination of calibration; Q2_C; Q2_CV: coefficient of determination of cross-validation; Q2_T coefficient of determination of test set**

Metrics	VCPA-IRIV	BOSS	BOSS-IRVS with three variables added	BOSS-IRVS with six variables added
mnLV	9	9	9	9
nVAR	35±5	27.7±8.2	30	33
RMSEC	0.4830± 0.0099	0.4829±0.0074	0.4697± 0.0011	0.4624±7.8510e-04
RMSEV	0.5210±0.0110	0.5341±0.0063	-	-
RMSEP	0.6004±0.0075	0.6366±0.0137	0.6026±0.0025	0.5965±0.0039
Q2_C	0.9946±2.2305e-04	0.9946±1.6777e-04	0.9949±2.3538e-05	0.9950±1.6897e-05
Q2_CV	0.9937±2.6812e-04	0.9934±1.5671e-04	-	-
Q2_T	0.9901±2.4777e-04	0.9889±4.7175e-04	0.9901±8.1650e-05	0.9903±1.2722e-04

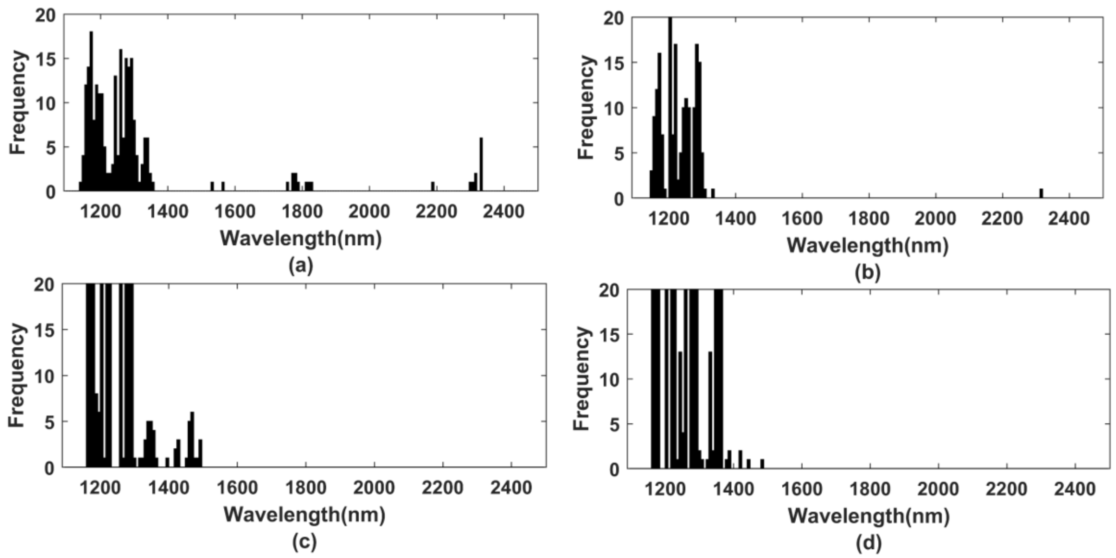


FIGURE 11. **The frequency of selected variables within 20 times on the wheat protein dataset: (A) VCPA-IRIV (B) BOSS, (C) proposed method selected three variables (D) proposed method selected six variables.**

D. wheat protein dataset

From Table 5 and Figure 11, BOSS-IRVS can be seen clearly achieving better results compared with VCPA-IRIV and BOSS. The values of RMSEP and Q2_Tof the BOSS-IRVS are, respectively, 0.1789 and 0.9119 compared to 0.2089 and 0.8779 for BOSS, and 0.2235 and 0.8602 for VCPA-IRIV. The BOSS-IRVS with three added variables outperforms the BOSS-IRVS with six added variables which the reason for overfitting. The value of RMSEC for BOSS-IRVS with six added variables has low RMSEC and high RMSEP compared with the BOSS-IRVS with three added variables.

The variables around 1104-1400nm can be selected by all methods which indicate the importance of this region which corresponds to the first overtone of the O-H stretch bond vibration [7]. The VCPA-IRIV select other variables in intervals around 1800 and between 2200 and 2400. The BOSS and the BOSS-IRVS concentrated on this region which shows better performance than VCPA-IRIV. The proposed method combines the variables select by BOSS and add only three variables selected on the important intervals and showed a significant improvement of the prediction accuracy—the lowest variables selected by BOSS followed by VCPA-IRIV and BOSS-IRVS method. Table 5 and Table 1 showed the proposed hybrid method and previous different variable selection methods on the wheat protein dataset. We analyzed that, MC-UVE and CARS method select informative variable; however, it selects another variable in uninformative intervals. IVSO and GA-PLS-LRC method select informative intervals and concentrate their variables in these informative variables which lead to a good performance. IVSO outperformed PLS, CARS and MC-UVE while GA-PLS-LRC outperformed GA-PLS. A recent paper called SMCPA showed a good concentration with a low number of variables and outperformed BOSS, VCPA, and CARS. Our proposed hybrid method proved that adding three informative variables in informative intervals could improve the result significantly.

E. hemoglobin dataset

From Table 6, it is clear that the BOSS-IRVS can achieve better results compared with VCPA-IRIV and BOSS in terms of RMSEP and Q2_T. The values of the RMSEP are 0.4114, 0.4270, 0.4368, and 0.5167 for BOSS-IRVS when six added variables, the BOSS-IRVS when three added variables, VCPA-IRIV and BOSS respectively. The values of Q2_Test are 0.9788, 0.9772, 0.9760, and 0.9663 for BOSS-IRVS for six added variables, BOSS-IRVS for three added variables, VCPA-IRIV, and BOSS methods respectively. Figure 12 showed that the intervals between 1600 and 1800, and between 2200 and 2400 are select by all methods. VCPA-IRIV and BOSS selected between 1200 and 1400. The BOSS-IRVS method added six variables only in intervals between 1600 and 1800 and between 2200 and 2400. The selection of these variables improved the result of the BOSS method significantly and outperformed VCPA-IRIV. All three methods select two intervals indicating the importance of these intervals. In addition, the BOSS-IRVS select fewer variables compared to VCPA-IRIV. However, the BOSS-IRVS outperformed the VCPA-IRV, the percentage of improvement for the hemoglobin dataset is 5.8 for VCPA-IRIV. Besides, when only six variables added to the BOSS, the result improved significantly to 20.3 %.

TABLE 5. **Results for the wheat protein dataset. nVAR: number of variables; mnLVs: max number of latent variables; RMSEC: root mean-square error of calibration RMSEV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; coefficient of determination of calibration; Q2_C; Q2_CV: coefficient of determination of cross-validation; Q2_T coefficient of determination of test set**

Metrics	VCPA-IRIV	BOSS	BOSS-IRVS with three variables added	BOSS-IRVS with six variables added
mnLV	10	10	10	10
nVAR	11.9± 2.9	9±1.3	13	16
RMSEC	0.2815±0.0106	0.2789±0.0050	0.2697±0.0022	0.2496±4.6216e-04
RMSEV	0.3146±0.0122	0.3071±0.0045	-	-
RMSEP	0.2235±0.0314	0.2089±0.0285	0.1789±0.0050	0.1868±0.0048
Q2_C	0.9447±0.0042	0.9457±0.0019	0.9493±8.3443e-04	0.9566± 1.6081e-04
Q2_CV	0.9309±0.0053	0.9342±0.0019	-	-
Q2_T	0.8602± 0.0403	0.8779±0.0363	0.9119±0.0048	0.9041± 0.0050

TABLE 6. **Results for hemoglobin dataset. nVAR: number of variables; mnLVs: max number of latent variables; RMSEC: root mean-square error of calibration RMSEV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; coefficient of determination of calibration; Q2_C; Q2_CV: coefficient of determination of cross-validation; Q2_T coefficient of determination of test set.**

Metrics	VCPA-IRIV	BOSS	BOSS-IRVS with three variables added	BOSS-IRVS with six variables added
mnLV	10	10	10	10
nVAR	28.5±6.7	20.2±5.6	17	20
RMSEC	0.2128±0.0144	0.2668±0.0053	0.2530±0.0014	0.2446±0.0015
RMSEV	0.2278±0.0142	0.2876±0.0059	-	-
RMSEP	0.4368±0.0364	0.5167±0.0474	0.4270±0.0072	0.4114±0.0045
Q2_C	0.9844±0.0022	0.9755±9.6038e-04	0.9780±2.3960e-04	0.9794±2.4671e-04
Q2_CV	0.9821±0.0023	0.9715±0.0012	-	-
Q2_T	0.9760±0.0040	0.9663±0.0059	0.9772±7.6934e-04	0.9788±4.6791e-04

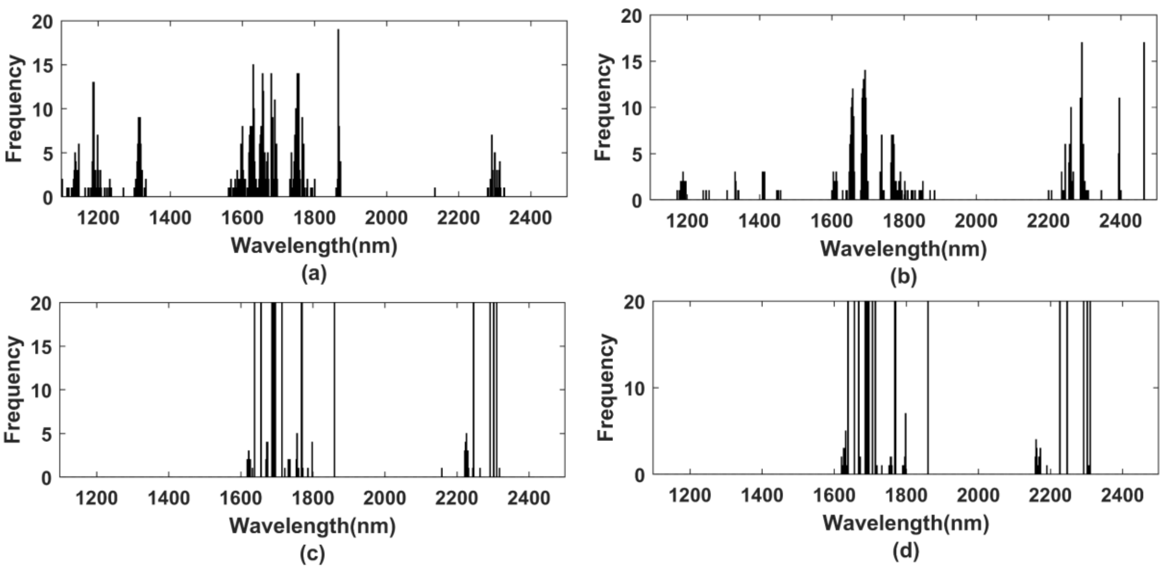


FIGURE 12. **The frequency of selected variables within 20 times on the hemoglobin dataset: (A) VCPA-IRIV (B) BOSS, (C) proposed method selected three variables (D) proposed method selected six variables.**

VI. CONCLUSIONS AND FUTURE WORKS

To conclude, a new hybrid strategy for variable selection has been proposed (BOSS-IRVS) in this study. The hybrid strategy takes full advantage of BOSS as proved to select informative intervals and uses interval random variables selection to search informative variables in the informative interval selected by BOSS. It solves the problem of BOSS's tendency to select fewer variables, and also improve the predictive accuracy. Seven NIR datasets were used to investigate the improvement of this hybrid strategy. The results show that the hybrid strategy significantly improved the model's prediction performance when compared with two high-performance methods (BOSS and VCPA-IRIV). It is worth pointing out that the proposed hybrid strategy is general and can be coupled with some other optimization or variable selection methods for further optimization. Although it was employed on the kind of NIR dataset in this study, it could be applied to other kinds of high dimensional data, such as genomics, proteomics, metabolomics, QSAR, and others. In future work, we will consider applying our proposed model in high performance variable selection method such as FOSS, SOBSS and SMCPA. Besides, we will consider the computational cost in the performance evaluation.

REFERENCES

[1] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, and J. Pagès, "NIR calibration in non-linear systems: Different PLS approaches and artificial neural networks," *Chemom. Intell. Lab. Syst.*, vol. 50, no. 1, pp. 75–82, 2000, doi: 10.1016/S0169-7439(99)00048-9.

[2] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009, doi: 10.1016/j.trac.2009.07.007.

[3] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003, doi: 10.1162/153244303322753616.

[4] B.-C. Deng *et al.*, "A bootstrapping soft shrinkage approach for variable selection in chemical modeling," *Anal. Chim. Acta*, vol. 908, pp. 63–74, Feb. 2016, doi: 10.1016/J.ACA.2016.01.001.

[5] H. Yan *et al.*, "A modification of the bootstrapping soft shrinkage approach for spectral variable selection in the issue of over-fitting, model accuracy and variable selection credibility," *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 210, pp. 362–371, 2019, doi: 10.1016/j.saa.2018.10.034.

[6] Y. W. Lin, B. C. Deng, L. L. Wang, Q. S. Xu, L. Liu, and Y. Z. Liang, "Fisher optimal subspace shrinkage for block variable selection with applications to NIR spectroscopic analysis," *Chemom. Intell. Lab. Syst.*, vol. 159, no. November, pp. 196–204, 2016, doi: 10.1016/j.chemolab.2016.11.002.

[7] Y. Wang, Z. Jia, and J. Yang, "An Variable Selection Method of the Significance Multivariate Correlation Competitive Population Analysis for Near-Infrared Spectroscopy in Chemical Modeling," *IEEE Access*, vol. 7, pp. 167195–167209, 2019, doi: 10.1109/ACCESS.2019.2954115.

[8] J. H. Jiang, R. James, B. H. W. Siesler, and Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data," *Anal. Chem.*, vol. 74, no. 14, pp. 3555–3565, 2002, doi: 10.1021/ac011177u.

[9] Y. Yun *et al.*, "A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration," *Anal. Chim. Acta*, vol. 1058, pp. 58–69, 2019, doi: 10.1016/j.aca.2019.01.022.

[10] G. Tang *et al.*, "A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm," *Analyst*, vol. 139, no. 19, pp. 4894–4902, 2014, doi: 10.1039/c4an00837e.

[11] H. Chen, X. Liu, Z. Jia, Z. Liu, K. Shi, and K. Cai, "A combination strategy of random forest and back propagation network for variable selection in spectral calibration," *Chemom. Intell. Lab. Syst.*, vol. 182, pp. 101–108, 2018, doi: 10.1016/j.chemolab.2018.09.002.

[12] Y. Yun, H. Li, B. Deng, and D. Cao, "An overview of variable selection methods in multivariate analysis of near-infrared spectra," *Trends Anal. Chem.*, 2019, doi: 10.1016/j.trac.2019.01.018.

[13] L. L. Wang *et al.*, "A selective review and comparison for interval variable selection in spectroscopic modeling," *Chemom. Intell. Lab. Syst.*, vol. 172, no. September 2017, pp. 229–240, 2018, doi: 10.1016/j.chemolab.2017.11.008.

[14] H. D. Li, Q. S. Xu, and Y. Z. Liang, "Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification," *Anal. Chim. Acta*, vol. 740, pp. 20–26, 2012, doi: 10.1016/j.aca.2012.06.031.

[15] Y.-H. Yun *et al.*, "A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration," *Anal. Chim. Acta*, vol. 807, pp. 36–43, Jan. 2014, doi: 10.1016/J.ACA.2013.11.032.

[16] B. Deng, Y. Yun, Y. Liang, and L. Yi, "A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling," *Analyst*, vol. 139, no. 19, p. 4836, Jul. 2014, doi: 10.1039/C4AN00730A.

[17] W. Wang, Y. Yun, B. Deng, W. Fan, and Y. Liang, "Iteratively variable subset optimization for multivariate calibration," *RSC Adv.*, vol. 5, no. 116, pp. 95771–95780, 2015, doi: 10.1039/c5ra08455e.

[18] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Anal. Chim. Acta*, vol. 648, no. 1, pp. 77–84, Aug. 2009, doi: 10.1016/J.ACA.2009.06.046.

[19] K. Zheng *et al.*, "Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra," *Chemom. Intell. Lab. Syst.*, vol. 112, pp. 48–54, 2012, doi: 10.1016/j.chemolab.2012.01.002.

[20] R. Zhang *et al.*, "A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection," *Chemom. Intell. Lab. Syst.*, vol. 175, no. January, pp. 47–54, 2018, doi: 10.1016/j.chemolab.2018.02.007.

[21] Y. H. Yun *et al.*, "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration," *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 111, pp. 31–36, 2013, doi: 10.1016/j.saa.2013.03.083.

[22] B. C. Deng, Y. H. Yun, P. Ma, C. C. Lin, D. B. Ren, and Y. Z. Liang, "A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals," *Analyst*, vol. 140, no. 6, pp. 1876–1885, 2015, doi: 10.1039/c4an02123a.

[23] X. Song, Y. Huang, H. Yan, Y. Xiong, and S. Min, "A novel algorithm for spectral interval combination optimization," *Anal. Chim. Acta*, vol. 948, pp. 19–29, 2016, doi: 10.1016/j.aca.2016.10.041.

[24] S. Ye, D. Wang, and S. Min, "Successive projections algorithm combined with uninformative variable elimination for spectral variable selection," *Chemom. Intell. Lab. Syst.*, vol. 91, no. 2, pp. 194–199, 2008, doi: 10.1016/j.chemolab.2007.11.005.

[25] X. Fu, F. J. Duan, T. T. Huang, L. Ma, J. J. Jiang, and Y. C. Li, "A fast variable selection method for quantitative analysis of soils using laser-induced breakdown spectroscopy," *J. Anal. At. Spectrom.*, vol. 32, no. 6, pp. 1166–1176, 2017, doi: 10.1039/c7ja00114b.

[26] B.-C. Deng, Y.-H. Yun, and Y.-Z. Liang, "Model population analysis in chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 149, pp. 166–176, Dec. 2015, doi: 10.1016/J.CHEMOLAB.2015.08.018.

[27] L. P. Brá, M. Lopes, A. P. Ferreira, and J. C. Menezes, "A bootstrap-based strategy for spectral interval selection in PLS regression," *J. Chemom.*, vol. 22, no. 11–12, pp. 695–700, 2008, doi: 10.1002/cem.1153.

[28] H. D. Li, Y. Z. Liang, Q. S. Xu, and D. S. Cao, "Model population analysis for variable selection," *J. Chemom.*, vol. 24, no. 7–8, pp. 418–423, 2010, doi: 10.1002/cem.1300.

[29] H. Ali Gamal Al-Kaf, A. M. Mohsen, and K. Seng Chia, "Improved model population analysis in near infrared spectroscopy," *First Int. Conf. Intell. Comput. Eng.*, pp. 1–9, 2019, doi: 10.1109/ICOICE48418.2019.9035177.

[30] R. Zhang *et al.*, "A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection," *Chemom. Intell. Lab. Syst.*, vol. 175, pp. 47–54, Apr. 2018, doi: 10.1016/J.CHEMOLAB.2018.02.007.

[31] Y.-H. Yun *et al.*, "A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration," *Anal. Chim. Acta*, vol. 807, pp. 36–43, Jan. 2014, doi: 10.1016/J.ACA.2013.11.032.

[32] Y. H. Yun *et al.*, "A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration," *Chemom. Intell. Lab. Syst.*, vol. 130, pp. 76–83, 2014, doi: 10.1016/j.chemolab.2013.09.007.

[33] M. Forina *et al.*, "Transfer of calibration function in near-infrared spectroscopy," *Chemom. Intell. Lab. Syst.*, vol. 27, no. 2, pp. 189–203, 1995, doi: 10.1016/0169-7439(95)80023-3.

[34] J. H. Kalivas, "Chemometricsand intelligent laboratory systems Two data sets of near infrared spectra," *Chemom. Intell. Lab. Syst.*, vol. 37, pp. 255–259, 1997.

[35] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007, doi: 10.1214/009053606000001523.

[36] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[37] R. Leardi, "Application of genetic algorithm-PLS for feature selection in spectral data sets," *J. Chemom.*, vol. 14, no. 5–6, pp. 643–655, 2000, doi: 10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E.

[38] B. Igne *et al.*, "The 2010 IDRC Software Shoot-out at a Glance," *NIR news*, vol. 21, no. 8, pp. 14–16, Dec. 2010, doi: 10.1255/nirn.1216.

[39] M. N. E. M. Idrus, K. S. Chia, H. M. Sim, and H. A. G. Al-kaf, "Artificial neural network and Savitzky Golay derivative in predicting blood hemoglobin using near-infrared spectrum," *Int. J. Integr. Eng.*, vol. 10, no. 8, pp. 112–119, 2018, doi: 10.30880/ijie.2018.10.08.017.