DIFFERENT HYBRID PREDICTION'S MACHINE LEARNING ALGORITHMS FOR QUANTITATIVE ANALYSIS IN LASER-INDUCED BREAKDOWN SPECTROSCOPY

Mohsen Rezaei [1], Fatemeh Rezaei [2*], Parvin Karimi [3]

[1] Department of Industrial Engineering, University of Science and Technology of Mazandaran, Behshahr, Iran
[2] Department of Physics, K. N. Toosi University of Technology, Tehran, Iran; e-mail: fatemehrezaei@kntu.ac.ir
[3] Department of Physics, South Tehran Branch, Islamic Azad University, Tehran, Iran

Laser-induced breakdown spectroscopy (LIBS) technique is employed for quantitative analysis of aluminum samples by different classical machine learning approaches. A Q-switch Nd:YAG laser at a fundamental harmonic of 1064 nm is utilized for the creation of LIBS plasma in order to predict constituent concentrations of the aluminum standard alloys. In the current research, concentration prediction is performed by linear approaches of support vector regression (SVR), multiple linear regression (MLR), principal component analysis (PCA) integrated with MLR (PCA–MLR), and SVR (PCA–SVR), as well as nonlinear algorithms of artificial neural network (ANN), kernelized support vector regression (KSVR), and the integration of traditional principal component analysis with KSVR (PCA–KSVR), and ANN (PCA–ANN). Furthermore, dimension reduction is applied to various methodologies by the PCA algorithm in order to improve the quantitative analysis. The results indicated that the combination of PCA with the KSVR algorithm model had the best efficiency in predicting most of the elements among other classical machine learning algorithms.

Keywords: *laser-induced breakdown spectroscopy, classical machine learning algorithms, principal component analysis, concentration prediction, quantitative analysis.*

ГИБРИДНЫЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ КОЛИЧЕСТВЕННОГО СОСТАВА АЛЮМИНИЕВЫХ СПЛАВОВ МЕТОДОМ ЛАЗЕРНО-ИСКРОВОЙ ЭМИССИОННОЙ СПЕКТРОСКОПИИ

M. Rezaei [1], F. Rezaei [2*], P. Karimi [3]

УДК 543.423

[1] Университет науки и технологий Мазандарана, Бехшахр, Иран
[2] Технологический университет K. N. Toosi, Тегеран, Иран; e-mail: fatemehrezaei@kntu.ac.ir
[3] Южно-Тегеранский филиал Исламского университета Азад, Тегеран, Иран

(Поступила 25 мая 2022)

Проведен количественный анализ образцов алюминия с помощью лазерно-искровой эмиссионной спектроскопии (LIBS) и различных классических подходов к машинному обучению. Nd:YAG-лазер с модулятором добротности на основной гармонике 1064 нм используется для создания плазмы LIBS для прогнозирования концентраций составляющих стандартных алюминиевых сплавов. Прогнозирование концентрации выполнено с помощью линейных подходов регрессии опорных векторов (SVR), множественной линейной регрессии (MLR), анализа главных компонент (PCA), интегрированного с MLR (PCA–MLR) и SVR (PCA–SVR), а также нелинейных алгоритмов искусственной нейронной сети (ANN), ядерной регрессии опорных векторов (KSVR) и объединения традиционного анализа главных компонент с KSVR (PCA–KSVR) и ANN (PCA–ANN). Для улучшения количественного анализа применяется уменьшение размерности к различным методологиям на основе алгоритма PCA. Ком-

бинация PCA–KSVR показывает наибольшую эффективность в прогнозировании большинства элементов среди других классических алгоритмов машинного обучения.

Ключевые слова: *лазерно-искровая эмиссионная спектроскопия, классические алгоритмы машинного обучения, анализ главных компонент, прогнозирование концентрации, количественный анализ.*

Introduction. Laser-induced breakdown spectroscopy (LIBS) is an analytical technique that correlates the spectral signal to the concentration of the analyte according to different mathematical calculations [1]. LIBS is a simple analytical method to identify the elemental composition which uses a focused high-energy laser pulse for the generation of plasma from the solid, liquid, or gaseous samples. In recent decades, multivariate classical machine learning algorithms, as the new methodologies, have attracted a lot of interest for quantitative analysis in LIBS spectroscopy. LIBS technique, as a powerful method, is an online and fast kind of atomic emission spectroscopy for concentration prediction. Here, the spectroscopic analysis of the light emitted by the laser-produced plasma is used for the identification of the constituent elements of the analyzed sample [2–4].

Different research groups have investigated multivariate studies in LIBS spectroscopy by different techniques of artificial neural networks (ANN) algorithm [5–9], principal component analysis method (PCA) [10–12], support vector regression (SVR) [13–15], and multiple linear regression (MLR) technique [16–18]. For instance, Unnikrishnan et al [19] have employed PCA for the classification of four widely used plastics in LIBS spectroscopy. They have shown that the 375–390 nm region of the LIBS spectra illustrated good results in comparison to other regions without much of the preprocessing. In addition, Ferreira et al. [20] have used an artificial neural network for calibration strategy in the LIBS technique, aiming to achieve Cu determination in soil samples. They have presented adequate LOD by utilizing a portable LIBS instrument. Moreover, Dong et al. [21] have explored the carbon contents in coal samples by LIBS by MLR algorithms, the partial least squares regression (PLSR), and SVR. They have illustrated that the combination of carbon atomic and molecular spectra with both PLSR and SVR correction improved the quantitative analysis, and the SVR correction helped in reaching better measurement accuracy.

This study represents a combination of LIBS emission spectra with different prediction models. In the current work, the focus is on the improvement of the precision of the quantitative analysis in LIBS spectroscopy by introducing the best multivariate methodology. Here, a comparison is made among eight multivariate algorithms of MLR, SVR, kernelized support vector regression (KSVR), ANN, PCA–MLR, PCA-SVR, PCA–KSVR, and PCA–ANN in terms of both accuracy and precision based on LIBS. To the authors' best knowledge, the combination of the PCA model with other statistical methods is presented for the first time in LIBS spectroscopy for prediction purposes, resulting in impressive results. All of these methods are used to quantify the corresponding components of Si, Fe, Cu, Zn, Mn, and Mg in seven aluminum standard samples. Mean squared error (MSE), and the mean absolute error (MAE) are employed to evaluate the prediction ability of the mentioned statistical models which are indications of the prediction's concentrations.

Experimental. A typically utilized experimental setup of LIBS spectroscopy is presented in Fig. 1 [22, 23]. Here, a *Q*-switched Nd:YAG laser at a fundamental harmonic of 1064 nm wavelength, and 10 ns pulse duration, with a repetition rate of 10 Hz, and a laser energy of 50 mJ is used for plasma creation. The samples are different aluminum standards (1100 series) supplied by the Razi metallurgical research center in Iran. At a particular delay time, the laser light is conducted to a beam splitter and divided into two sections. One part is guided to the photodiode for launching the delay generator. Then, the ICCD camera gets a pulse from the delay generator to begin the acquisition of data. The other part is passed through a λ/2 plate and a Glan–Taylor prism to change the laser energy. The focusing of the laser pulse is performed by a lens with a 20-cm focal length. Moreover, the strike position of the laser pulse is adjusted by an *XYZ* stage throughout the experiment. Spatially integrated plasma emissions are collected, utilizing a quartz lens with the help of an objective lens, and then sent to an optical fiber. At the next stage, plasma radiations are guided to an Echelle spectrograph (Kestrel, SE200) to receive a spectrally resolved light spectrum with spectral ranges of 200–900 nm and a spectral resolution of 0.02 nm. During the adjustment of the gate and the delay time of the ICCD camera (Andor, iStar DH734), the recorded spectral emissions can be temporally studied. For the spectral analysis, the acquisition delay time between the laser pulse and the beginning of the acquisition is changed. Then, the optimum delay time is selected for experimental analysis in order to maximize the signal-to-background ratio (SNR) of the spectral line. In the current research, for each sample, 87 LIBS spectra are extracted for each irradiated spot which is repeated 10 times and averaged (in total, 870 spectra per sample) to predict constituent concentrations of the test aluminum standard alloys.

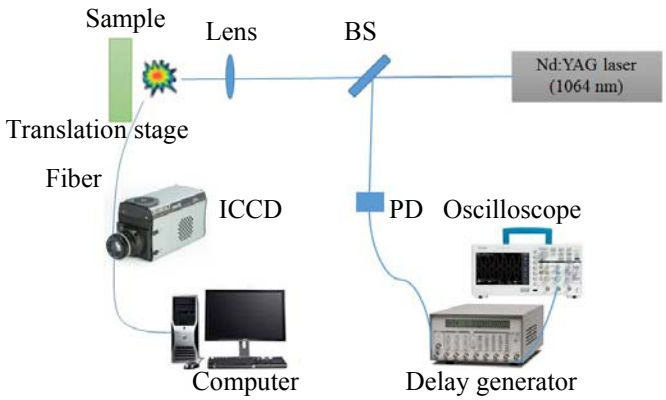

Fig. 1. Schematic of LIBS experimental set-up.

Statistical prediction methodologies. There are different machine learning algorithms to predict unknown quantities. In this research, the proposed methodologies employed 870 spectra of different standard aluminum samples in correspondence to different concentrations of Si, Fe, Cu, Al, Zn, Mn, and Mg. It should be mentioned that due to the high concentration of aluminum elements in all of the aluminum standard samples which induce self-absorption phenomenon, they are not utilized for prediction. It can be noted that all the chemometrics codes explained in this section including MLR, ANN, SVR, KSVR, and integrated PCA are calculated by MATLAB 2019b (MathWorks). The details of these algorithms are explained in the following sections.

Multiple linear regression method is a statistical multivariate technique that establishes a correlation between the independent and dependent variables (or criteria). MLR is an extensively coupled chemometrics method with LIBS spectroscopy which follows a linear predictor function as shown in the following [21]:

$$y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_N X_N + \varepsilon, \quad (1)$$

where a_0, a_1, a_N are the regression coefficients, ε is the residual error, and y is the dependent variable in a way that this equation must be written for all M samples.

In the current research, for designing the MLR model, spectral intensities at different concentrations are considered as independent variables, while elemental concentrations are regarded as dependent parameters. MATLAB software is utilized to calculate the regression equation and analyze of the results.

Support vector regression methodology is a kernel-based regression method which acts according to the principle of a support vector machine (SVM). The SVR method is introduced as a powerful algorithm for function estimation and pattern recognition. Here, n points with coordinates of (x_1,y_1), …, (x_n,y_n), where x_i denotes the input spectrum, and y_i indicates the intensities that correspond to the target value, are used as training dataset in which n is related to the number of samples. In the SVM algorithm, the hyperplane acts as a separating line between two data sets, but in the SVR method, the line is exploited to predict the continuous output. In this case, the margin is a region bounded between two hyperplanes. The main goal is the reduction of error, individualizing the hyperplane which maximizes the margin by taking into account that part of the error is tolerated [24].

It can be emphasized that the SVR algorithm maps the dataset from the nonlinear low-dimensional space to linear high-dimensional with the application of the kernel function so that nonlinear data changes into linear data in a new coordinate. Hence, the SVR method changes the nonlinear relationship of input datasets by using different kernel functions. The main aim of SVR is finding a function $f(x)$ with a deviation from y_n, for each training point x, by a value not greater than ε. It should be stressed that each hyperplane can be written as the set of points x implying that $f(x)$ functions as flat as possible [25]:

$$f(x) = wx + b. \quad (2)$$

Here, b is the bias term and w is the normal vector to the hyperplane. Generally, SVR regression algorithm is formulated so that the following functional equation can be minimized as [26, 27]:

$$\frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}(\xi_n + \xi_n^*). \quad (3)$$

Subject to [27]:

$$y_n - (wx_n + b) \le \varepsilon + \xi_n,$$

$$(wx_n + b) - y_n \le \varepsilon + \xi_n^*, \quad \xi_n, \xi_n^* \ge 0, \tag{4}$$

where ξ_n and ξ_n^* are two positive slack variables for measuring the deviation; C is a box constraint constant with a positive value that controls many of the imposed on observations placed outside the epsilon margin (ε) and helps to prevent overfitting.

It should be mentioned that linear SVR in dual formula follows a Lagrangian function built from the primal function by regarding the nonnegative multipliers α_n and α^*_n for every observation x_n as:

$$y = \sum_{i=1}^{l} (\alpha_n - \alpha_n^*)(x_n x) + b\,. \tag{5}$$

Generally, SVR can be studied by a linear or nonlinear regression algorithm according to the kernel function used. It can be stressed that the Kernelized support vector regression (KSVR) method is faster than the standard SVR in the non-linear regression related to large datasets while producing the highest correctness in the prediction. SVR can support the nonlinear relationship of input datasets with different kernel functions [15].

In nonlinear SVR, the data will be transformed into a higher dimensional feature space by kernel functions ($K(x_n,x)$), for providing the linear separation as follows [28]:

$$y = \sum_{i=1}^{l} (\alpha_n - \alpha_n^*) K(x_n, x) + b\,. \tag{6}$$

Kernel functions can be explained by a polynomial function as:

$$K(x_n,x) = (x_n x)^d,\ d = 2, 3, \ldots, \tag{7}$$

or by Gaussian radial basis function as [26]:

$$K(x_n, x_j) = \exp\left(-\frac{\| x_n - x_j \|^2}{2\sigma^2}\right), \tag{8}$$

where σ is the width of the kernel function, and x_n, x_j are the nth and jth support vectors. It can be mentioned that in the linear form of SVR, the kernel function is equal to $x_n x$ in Eq. (6).

ANN, as an innovative computational approach, has been inspired right from its inception by recognizing that the human brain calculates in a completely different way from the conventional digital computer. Generally, the brain is a severely nonlinear information-processing system and parallel computer. It has the ability to manage its structural constituents (neurons) to analyze certain computations, such as perception and motor control, and pattern recognition much faster than the fastest digital computer. In fact, features of ANNs as a powerful mathematical technique can be considered by high parallelism, special data processing, acquiring knowledge through the learning process, nonlinear data mapping, highly weighted connections between elements, adaptability, and generalization ability. The most well-known applications of ANNs are indicated as classification, pattern recognition, and prediction in different areas, such as psychology or engineering [29]. These drastic simulation tools are comprised of an enormous number of single units such as artificial neurons, topology, and learning algorithms.

Actually, the neurons in ANN structures are simulated by transfer functions and are organized into input, output, and hidden layers. They are joined together with random coefficients (weights) which are continuously customized for performing optimization.

In the processing of information in a single node, first of all, the arrival of weighted activations accompanied by the previous nodes are combined together. Then, they pass through a transfer function as a feature of a specific node. After that, the outcome results created by all the nodes in one layer will be transmitted to the nodes of the next layer. This procedure continues to all of the layers for attaining the output data.

Figure 2 shows a schematic of a multilayer feed-forward network with some features of artificial neurons. It is constructed from three main sections: input layer with definite neurons equal to initial signals; hidden layers placed between the first and the last layers for the enhancement of responses of the specific arbitrary nodes; output layer consisting of the neurons which convey equivalent network results. It can be mentioned that none of the single-layer nodes are joined together and information can be transferred among layers according to the given weights. The estimation between the desired and the received results determines the efficiency of ANN. Actually, learning algorithms can be employed to obtain the most optimal responses with the least error.

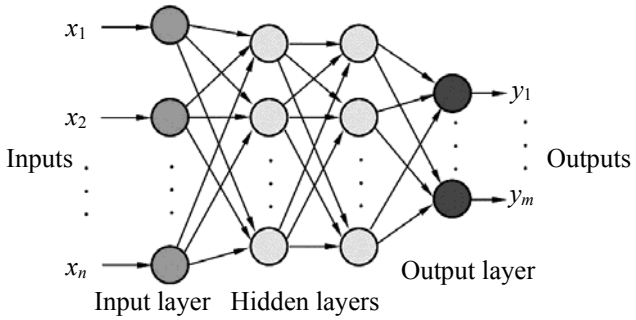

Fig. 2. A diagram of the structure of neural network [30].

Learning rules or training processes are explained as a group of input-output matching patterns that endorses obtaining the best results related to the input data. Consequently, the ANNs try to modify random weights as conjunction coefficients to increase the performance of networks by decreasing the occurrence of any error. The procedure depends on the forward and the backward propagation for modifying the value of all weights. It can be stressed that one of the most well-known learning instructions is named back-propagation (BP), which is based on the gradient-descent algorithm.

In the forward propagation, all of the connectivity components are constant and the final consequences are calculated. In the backward dispersion, all values of weights are refined based on the computed error in certain ways which then are used again for the subsequent forward processing.

PCA is a tool for the projection of a high-dimensional dataset into a small number of variables (named principal component, PC), and their utilization for the performance of a change of the data basis. PCA is a powerful chemometric methodology that is exploited to reduce data dimension by applying an orthogonal transformation. Here, two new matrices are generated: a loading matrix that shows the weight of the original variables, and a scoring matrix containing the projections of the samples so that x variables or raw data matrix are changed into a scoring matrix. It can be mentioned that a conversion is performed for the transformation of correlated variables into a group of uncorrelated variables (PCs). On the other hand, the PCs (score plot) are plotted and clusters are observed in the graph which is representative of the samples with similar composition/spectrum. The details of this method are given elsewhere in the [25]. In this study, PCA is done by using an in-house developed MATLAB routine (MathWorks INC., Natick, USA). The raw data matrix is constructed from spectral emissions of different aluminum samples [12].

Integration of PCA with other multivariate methods (PCA–ANN, PCA–SVR, PCA–KSVR, PCA–MLR). Generally, the integration of traditional principal component analysis with other multivariate methods can increase prediction accuracy. The strategic analysis is that PCA is used as a preprocessing step during another multivariate modeling for the reduction of the dimensionality of the original multivariable dataset. One of the main advantages of coupling PCA with other methodologies is the reduction of the training time in encountering large datasets processing since PCA makes data compression. Furthermore, for instance, ANN is very sensitive to correlations amongst inputs; therefore, in input data with strong correlations, PCA integration can remediate this problem. Consequently, for tackling these mentioned problems and observing the enhancement of the accuracy in a lot of prediction literature [31], this paper integrated PCA with ANN (called PCA–ANN), SVR (called PCA–SVR), KSVR (called PCA–KSVR), and MLR (called PCA–MLR) to find the best models.

Error estimation and accuracy evaluation. In statistical analysis, the concept of error is a fundamental concept for measuring the effectiveness of an estimator or predictor. It can be noted that different mathematical relations can be used to assess the errors. For instance, mean squared deviation (MSD) or mean squared error (MSE) estimates the quality of a predictor as follows [32]:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2. \tag{9}$$

In addition, root mean square error (RMSE) evaluates the goodness of the prediction for each trial as [32, 33]:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{\left(\hat{y}_i - y_i\right)^2}{n}}. \tag{10}$$

Mean absolute error (MAE) is also computed by considering the average of all absolute errors of the results as [34]:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \quad (11)$$

Here, y_i and $\hat{y}_i$ are the target and estimated concentration magnitudes corresponding to spectrum i, and n is the number of the test spectra taken into account.

Results and discussion. The most frequent methods for composition prediction in LIBS spectroscopy are calibration curve and artificial neural network approaches. In this work, quantitative determinations are carried out by using multivariate methods of linear approaches, such as MLR [35], PCA–MLR [36], SVR [15], and PCA–SVR [37], and nonlinear methods, such as ANN [38], KSVR [39], PCA–ANN [40], and PCA–KSVR [41].

A feed-forward perceptron ANN with a back-propagation algorithm is used for data prediction. In all of the calculations, MATLAB software is utilized and the ANN toolbox is employed for model development. Here, one sample is selected into the validation set for every seven samples, and the rest of the samples are used as the training set. It can be stressed that the data from five physical samples are used for training and only one physical sample is used for validation. This has been repeated by randomly selecting one validation sample from the six samples, carrying out six iterations in total. After optimizing each regression model in this way, the last (seventh) sample has been used for testing of model. The Levenberg-Marquardt algorithm, which is the most widely used learning algorithm is applied in this research for training the network [42].

Indeed, it can be stressed that based on different works of literature [29, 43], increasing the number of hidden layers induces the "overtraining" problem; therefore, it is seldom necessary to consider more than one hidden layer. The performance of the ANN model is shown in Table 1 for a typical Fe element versus different numbers of hidden neurons. As shown in this table, the best ANN model is obtained with just one neuron in the hidden layer which is highlighted with red color. It can be noted that this process can be repeated in a similar way for other elements.

TABLE 1. Determination of the Number of the Used Hidden Neurons for the Fe Element

Number of hidden neurons	MSE	MAE
1	0.016	0.093
2	0.021	0.110
3	0.029	0.113
4	0.033	0.149
5	0.040	0.159
6	0.040	0.163
7	0.033	0.148
8	0.066	0.215
9	0.090	0.242
10	0.077	0.213

In ANN algorithms, the trial-and-error method is employed for choosing the best transfer functions. Here, different transfer functions of linear (purelin), Log-sigmoid (logsig), and tangent sigmoid (tansig) [44, 45] are tested for all of the elements in order to attain the best prediction with minimum errors.

It is well known that the observed errors in the neural network analysis are dependent on the initial random magnitudes of the neural weights. It can be stressed that in this research, the calculations are repeated 30 times for each network and the error of the network is actually reported as the average of MSEs, and MAEs over all of the executions. A summary of the best transfer function for both hidden and output layers with an optimum number of hidden neurons is presented in Table 2. Furthermore, in this table, the relative error bars of MSE and MAE are illustrated for showing the precision of measurements. As can be seen, the results demonstrate that the best performance of the proposed ANN method has happened with different transfer functions related to the hidden and output layers for each element. Additionally, it is observed that Fe has the least error values among other elements during the usage of purely transfer functions in both of the layers. This fact can be attributed to the low concentration of the Fe which follows a linear trend in the calibration curve.

TABLE 2. The Best Transfer Functions of the Hidden and Output Layers and the Optimum Number of Hidden Neurons Based on ANN Calculations

Elements	Number of hidden neurons	Transfer functions		Errors	
		Hidden layer	Output layer	MSE	MAE
Fe	1	purelin	purelin	0.016	0.093
Zn	10	logsig	tansig	0.084	0.288
Si	1	purelin	purelin	0.142	0.306
Mn	1	purelin	purelin	0.063	0.207
Cu	3	tansig	purelin	0.033	0.148
Mg	6	logsig	logsig	1.378	1.174

N o t e. Error values of MSE and MAE represent the precision of the utilized methods.

Figure 3 presents the results of the best predictions for different elements of Fe, Cu, Zn, Si, and Mn with the ANN method. As it is clearly seen, the slope of the correlation curve between the predicted and the nominal concentration is near one for all of the elements.

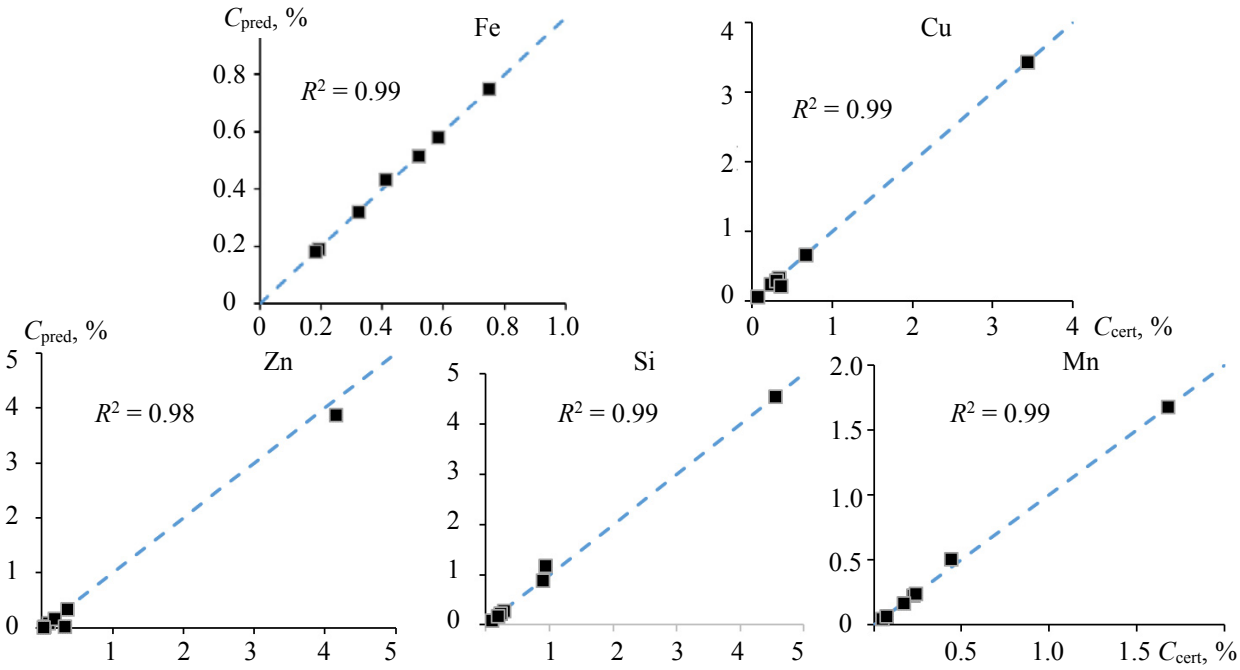

Fig. 3. Concentration prediction for different elements by artificial neural network.

The single SVR and MLR models simply utilize these methods to input variables for fore-casting the element concentration without using any PCA, or other preprocessing tools, such as linear independent component analysis (ICA) or nonlinear ICA. A comparison of the analytical performances of MLR and SVR methods is shown in Fig. 4 for the Si element due to its best prediction. Si concentration shows the best prediction by both SVR and MLR models with a low value for MSE and MAE. Here, the prediction models of MLR and SVR approaches as two linear methods are introduced as quantitative models which use the composition and spectral intensities of aluminum alloys for developing their performance. The correlation between the certified and predicted concentrations of these two models in this research represents a good quantitative measure of their prediction. It is obvious that the analytical predictions of the two models are very similar for the Si element. Moreover, after calculations for other elements, it is clearly seen that in both approaches, sometimes great deviations are seen in high concentrations which is due to the self-absorption effects that cause the multivariate methods to fail to perform well in the prediction of aluminum's composition [46]. Besides, the comparison of the two methods of SVR and ANN shows that the SVR model is more

accurate than ANN in predicting aluminum contents except for Zn and Fe elements. Additionally, ANN forecasts more exact concentrations compared to the MLR approach except for the Si element.

Another approach for the prediction of the composition of the constructed aluminum alloys is using kernel-based KSVR. The influence of different kernel functions on composition prediction is illustrated in Table 5 and their performance is compared in terms of MSE and MAE. As it is seen in Table 3, the Gaussian kernel almost forecasts better results than their kernels functions in different elements, which is in good agreement with results [15, 47]. It is worth noting that the linear kernel produces the worst prediction performance for most of the elements. Additionally, it can be mentioned from the KSVR results that the error values decreased a lot in all elements which induce that integration with the kernel function improved the SVR analysis performance a lot. This fact proposes that KSVR, which comprises the kernels function, provides better performance than the single SVR method. Red colors in this table show the lowest values of errors for different elements. Consequently, it can be concluded that the KSVR method is able to reproduce the concentrations with acceptance of the standard errors. An example of the best prediction with the KSVR approach is depicted in Fig. 5 for Fe element with correlation of 0.97.

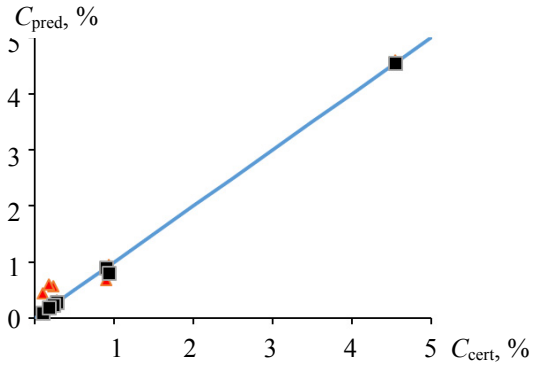

Fig. 4. A comparison between concentrations prediction by MLR (■) and SVR (▲) multivariate methods for Si element with coefficient of determination of $R^2 = 0.99$ and 0.96, respectively.

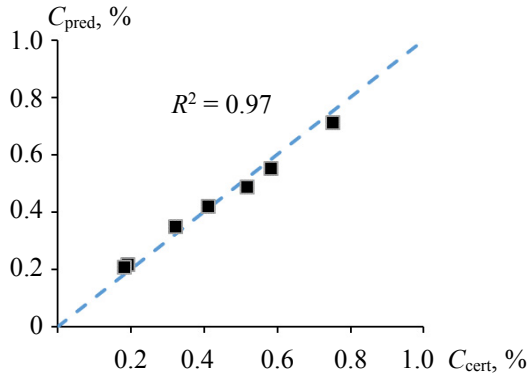

Fig. 5. Prediction of the concentration of the Fe element with the Kernelized support vector regression including correlation of 0.97.

TABLE 3. Comparison between Different Kernel Function's Performances for Different Aluminum Elements

Elements	Linear		Polynomial		Gaussian	
	MSE	MAE	MSE	MAE	MSE	MAE
Fe	0.078	0.280	0.0002	0.013	0.0001	0.012
Si	0.012	0.109	0.012	0.109	0.208	0.456
Mn	0.217	0.465	0.001	0.035	0.069	0.263
Cu	0.031	0.175	0.230	0.480	0.002	0.045
Mg	0.053	0.230	0.053	0.230	0.059	0.243
Zn	0.523	0.723	0.236	0.486	0.014	0.120

In the next stage, PCA algorithms are added to all of the mentioned methodologies including the ANN, MLR, SVR, and KSVR approaches. A feature which causes the application of the PCA integration methods in LIBS analysis to be much simpler (and more precise) as compared to other traditional algorithms, is reduction of dimension. In all assimilations with PCA methods, PCA is the first applied to the input variables to generate the PCs, then the considered analyses are conducted according to the generated PCs. Again, it can be mentioned that for representation of the characteristics of the input data, all PCs would be adopted to be used as new input variables for the mentioned model. More precisely, the proposed integrated approaches can be compared alone with their related method.

A comparison between the performance of the two methods of PCA–SVR and PCA–KSVR is presented in Fig. 6 for two elements of Fe and Mn. Approximately, similar trends are observed by both of the approaches.

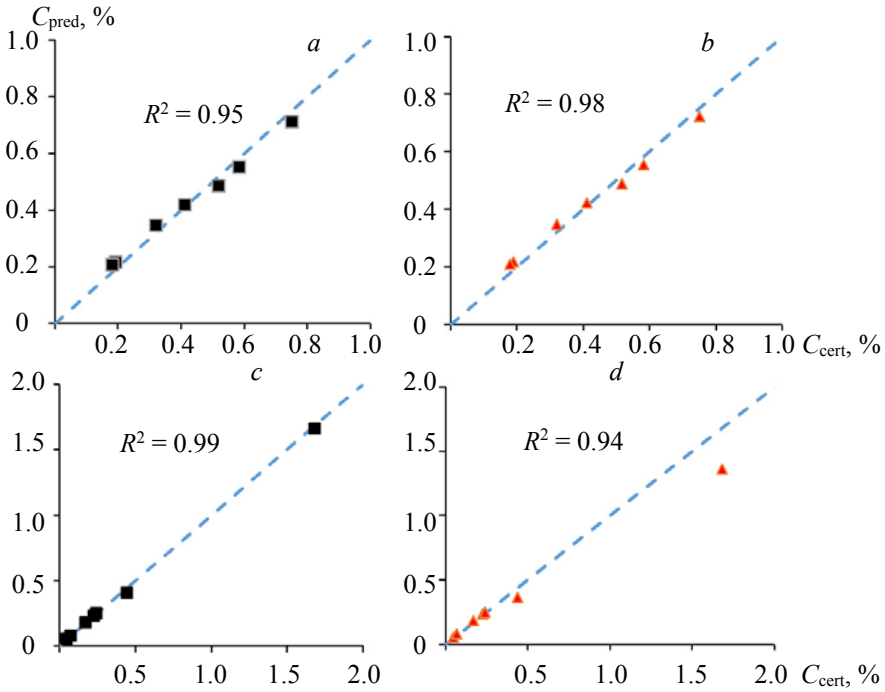

Fig. 6. Concentration predictions for two elements of Fe (a, b) and Mn (c, d) with combinational methods of PCA–SVR (b, d) and PCA–KSVR (a, c).

The performance of the combinational methods of PCA–ANN and PCA–MLR are compared for some elements including, Si, Mg, and Cu as shown in Fig. 7. As is seen in this figure, the accuracy of PCA–MLR is similar to PCA–ANN in the prediction of concentrations, while it is not true when the PCA technique is not combined with ANN and MLR methods. In this figure, the correlation slope of the curve between the predicted and the certified concentration is close to 1 for all of the elements.

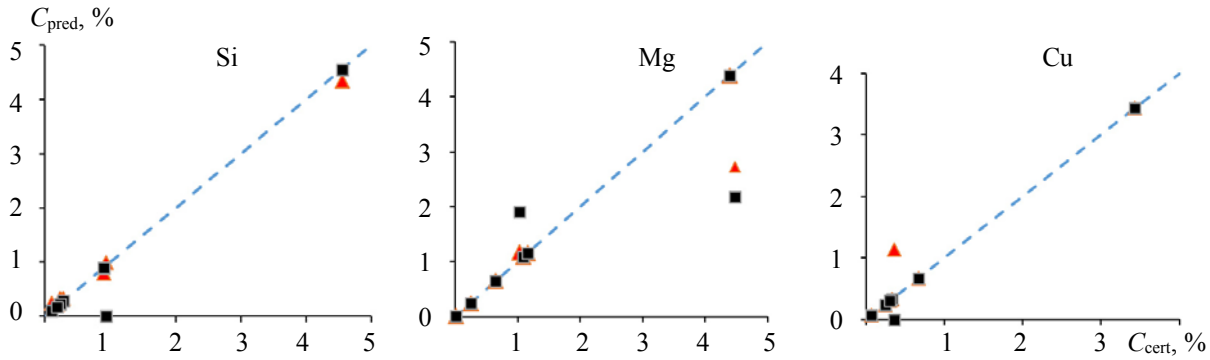

Fig. 7. Concentration prediction of Si, Mg, and Cu elements by two methods of PCA–ANN (▲) and PCA–MLR (■).

As a final consideration, a comparison summarized in Table 4 represents the prediction results of mean concentration by eight proposed approaches of ANN, MLR, SVR, KSVR, PCA–ANN, PCA–MLR, PCA–SVR, and PCA–KSVR. The validity of these methods is estimated by comparing the statistical error values of mean MSE and mean MAE calculated from the constituent elements of standard aluminum samples. It can be stressed that the least error values are highlighted with red color. As clearly seen in this table, in most of the elements but not all, the integration with PCA enhanced the accuracy of the single form of that special method. For instance, MLR methodology showed relatively high errors for the Mg, whereas applying PCA improved their prediction very much.

The results in Table 4 prove the evidence that the proposed PCA–KSVR as the best prediction methodology has produced the significantly lowest MAE, and RMSE (%) for all of the elements except for Si. From this table, it is found that the highest error value in the PCA–KSVR was achieved at 0.18% which reflects the robustness of this method. The comparison prediction results affirm that the PCA–KSVR approach not only improves the prediction accuracy of the single SVR method but also outperforms the competing methods in forecasting. Apparently, PCA–SVR can be introduced as the most accurate methodology after PCA–KSVR prediction method which produces low measurement errors (%) and high accuracy.

TABLE 4. Comparison between Different Multivariate Approaches in Prediction of Aluminum Contents by Calculation of Mean Squared Error (MSE), and the Mean Absolute Error (MAE) (%)

Method	Fe		Zn		Si		Mn		Cu		Mg	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MLR	0.020	0.140	0.326	0.571	0.015	0.121	1.518	1.232	1.935	1.391	4.066	2.016
PCA–MLR	0.079	0.280	1.447	1.203	0.865	0.930	0.087	0.295	0.130	0.360	0.768	0.877
SVR	0.029	0.171	2.413	1.553	**0.0004**	**0.020**	0.054	0.233	0.020	0.141	0.053	0.229
PCA–SVR	0.0002	0.012	**0.0004**	**0.021**	0.131	0.362	0.006	0.77	0.033	0.181	0.034	0.183
KSVR	0.0001	0.012	0.014	0.120	0.012	0.109	0.0012	0.035	0.0020	0.045	0.053	0.230
PCA–KSVR	**0.0001**	**0.012**	0.001	0.036	0.021	0.144	**0.0007**	**0.027**	**0.0017**	**0.041**	**0.033**	**0.183**
ANN	0.016	0.093	0.084	0.288	0.142	0.306	0.063	0.207	0.033	0.148	1.378	1.174
PCA–ANN	0.003	0.056	1.641	0.975	1.785	1.020	0.179	0.423	1.528	0.883	1.189	0.729

Generally, the integration of PCA and machine learning approaches illustrated good performance in different forecasting fields [27], such as the prediction of greenhouse gas emissions [48], evaluation of coronary artery diseases [49], and predicting the trend of the gasoline homogeneous charge compression ignition (HCCI) combustion behavior [50]. In Table 4, it is demonstrated that when the PC integration methods are not considered, the KSVR approach can better predict the compositions with higher precisions rather than simple MLR, SVR, or optimized artificial neural network models.

Taking into account the aforementioned considerations, the least MSE and the MAE of the prediction values for the validation samples are related to the Fe element and the worst ones are seen in the Mg element. This fact is due to the relatively low concentration of Fe and high content of Mg in aluminum alloys compared to other constituent elements; however, in the case of Mg, in order to improve the efficiency and reliability, PCA–KSVR meth-od can provide a great reduction in error calculations.

Finally, the best-proposed method is introduced in Table 5 for all of the elements along with a representative of the predicted values by these approaches. As shown, PCA–KSVR reported the nearest predicted values to certified concentrations for Fe, Zn, Mn, and Cu elements.

TABLE 5. Best Methods for Calculation of Concentrations of Different Elements of Fe, Zn, Si, Mn, Cu, and Mg

Element	Certified concentration	Predicted concentration	Best method
Fe	0.41	0.414	PCA–KSVR
Zn	0.32	0.341	PCA–SVR
Si	0.93	0.950	SVR
Mn	0.44	0.413	PCA–KSVR
Cu	0.36	0.401	PCA–KSVR
Mg	1.03	0.847	PCA–KSVR

Conclusions. The determination of the exact composition of the aluminum alloys is somewhat difficult by LIBS spectroscopy due to the important matrix effects and the nonlinear relation of the spectral intensities versus concentration. In this paper, different statistical prediction methods, such as MLR, ANN, SVR, KSVR, PCA–MLR, PCA–ANN, PCA–SVR, and PCA–KSVR, as methodological approaches are coupled with LIBS spectroscopy technique in order to evaluate the effectiveness of the proposed models and introduce the best quantitative methods. For each model, the general variation of the errors including MSE and MAE is reported. It was seen that in most of the cases, the assimilation of PCA significantly improved the

performance of the chosen approach as compared to the single formation of that model. The experimental results verified that PCA–KSVR is preferable with respect to other approaches only through the comparison of different error values. Finally, a definite benefit of the proposed approaches is the possibility of using them effectively for giving information on constituent elements of each arbitrary sample in LIBS analysis.

Acknowledgements. The authors are thankful to Professor Seyed Hassan Tavassoli for sharing his laboratory equipment and his kind support. The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. F. Rezaei, G. Cristoforetti, E. Tognoni, S. Legnaioli, V. Palleschi, A. Safi, *Spectrochim. Acta B: At. Spectrosc.*, **169**, 105878 (2020).
2. F. Rezaei, S. H. Tavassoli, *Phys. Plasmas*, **20**, 013301 (2013).
3. S. Messaoud Aberkane, A. Safi, A. Botto, B. Campanella, S. Legnaioli, F. Poggialini, S. Raneri, F. Rezaei, V. Palleschi, *Appl. Sci*., **10**, 4973 (2020).
4. F. Rezaei, *Appl. Opt.*, **59**, 3002 (2020).
5. F. Rezaei, P. Karimi, S. Tavassoli, *Appl. Phys. B*, **114**, 591 (2014).
6. Y .Yang, C. Li, S. Liu, H. Min, C. Yan, M. Yang, J. Yu, *Anal. Methods*, **10** (2020).
7. M. Boueri, V. Motto-Ros, W.-Q. Lei, L.-J. Zheng, H.-P. Zeng, *Appl. Spectrosc*., **65**, 307 (2011).
8. T. A. Labutin, A. M. Popov, S. N. Raikov, S. M. Zaytsev, N. A. Labutina, N. B. Zorov, *J. Appl. Spectrosc*., **80**, 315–318 (2013)
9. D. Luarte, A. K. Myakalwar, M. Velásquez, J. Álvarez, C. Sandoval, R. Fuentes, J. Yañez, D. Sbarbaro, *Anal. Methods*, **9** (2021)
10. G. S. Senesi, R. A. Romano, B. S. Marangoni, G. Nicolodelli, P. R. Villas-Boas, V. M. Benites, D. M. B. P. Milori, *J. Appl. Spectrosc.*, **84**, 923–928 (2017).
11. F. F. Hilario, M. L. de Mello, E. R. Pereira-Filho, *Anal. Methods*, **2** (2021).
12. S. Chatterjee, M. Singh, B. Biswal, U. K. Sinha, S. Patbhaje, A. Sarkar, *Anal. Bioanal. Chem.* **9**, 411 (2019).
13. Q. Q. Wang, L.A. He, Y. Zhao, Z. Peng, L. Liu, *Laser. Phys*., **26**, 065605 (2016).
14. T. F. Boucher, M. V. Ozanne, M. L. Carmosino, M. D. Dyar, S. Mahadevan, E. A. Breves, K. H. Lepore, S. M. Clegg, *Spectrochim. Acta B: At. Spectrosc*., **107**, 1 (2015).
15. J. Huang, M. Dong, S. Lu, W. Li, J. Lu, C. Liu, J.H. Yoo, *J. Anal. At. Spectrom.*, **33**, 720 (2018).
16. S. Laville, M. Sabsabi, F. R. Doucet, *Spectrochim. Acta B: At. Spectrosc.*, **62**, 1557 (2007).
17. J. Wang, S. Xue, P. Zheng, Y. Chen, R. Peng, *Anal. Lett*., **2017**, 50 (2000).
18. K. K. Ayyalasomayajula, V. Dikshit, F. Y. Yueh, J. P. Singh, L. T. Smith, *Anal. Bioanal. Chem.,* **400**, 3315 (2011).
19. V. K. Unnikrishnan, K. S. Choudhari, Suresh D. Kulkarni, R. Nayak, V. B. Kartha, C. Santhosh, *RSC Adv.*, **3**, 25872–25880 (2013).
20. E. C. Ferreira, D. M. B. P. Milori, E. J. Ferreira, R. M. Da Silva, L. Martin-Neto, *Spectrochim. Acta B: At. Spectrosc*., **63**, 1216–1220 (2008).
21. M. Dong, L. Wei, J. Lu, W. Li, S. Lu, S. Li, C. Liu, J. H. Yoo, *J. Anal. At. Spectrom*., **34**, 480–488 (2019).
22. A. Safi, S. H. Tavassoli, G. Cristoforetti, S. Legnaioli, V. Palleschi, F. Rezaei, E. Tognoni, *J. Adv. Res.*, **18**, 1 (2019).
23. H. Nozari, F. Rezaei, S. H. Tavassoli, *Phys. Plasmas*, **22**, 093302 (2015).
24. T. F. Boucher, M. V. Ozanne, M. L. Carmosino, M. D. Dyar, S. Mahadevan, E. A. Breves, K. Lepore, *Spectrochim. Acta B: At. Spectrosc.*, **107**, 1–10 (2015).
25. A. J. Smola, B. Schölkopf, *Statist. Comput.*, **14**, 199 (2004).
26. S. Thomas, G. N. Pillai, K. Pal, *Geomat. Nat. Hazards Risk*, **8**, 177 (2017).
27. S. Shokri, M. T. Sadeghi, M. A. Marvast, S. Narasimhan, *Chem. Ind. Chem. Eng. Q*, **21**, 379 (2015).
28. X. Ma, Y. Zhang, H. Cao, S. Zhang, Y. Zhou, *J. Spectrosc.* (**2018**).
29. M. Rezaei, S. Chaharsooghi, A. Kashan, R. Babazadeh, *Int. J. Environ. Sci. Technol*, **17**, 3241 (2020).
30. A. Pouliakis, E. Karakitsou, N. Margari, P. Bountris, M. Haritou, J. Panayiotides, D. Koutsouris, P. Karakitsos, *Biomed. Eng. Comput. Biol.*, **7** (2016).

31. A. Nikkhah, A. Rohani, K. A. Rosentrater, M. El Haj Assad, S. Ghnimi, *Environ. Prog. Sustain*, **38**, 13130 (2019).
32. E. C. Ferreira, J. A. G. Neto, D. M. B. P. Milori, E. J. Ferreira, J. M. Anzano, *Spectrochim. Acta B*: *At. Spectrosc.*, **110**, 96 (2015).
33. H. Qin, Z. Lu, S. Yao, Z. Liab, J. Lu, *JAAS*, **34**, 347 (2019).
34. E. D'Andrea, S. Pagnotta, E. Grifoni, G. Lorenzetti, S. Legnaioli, V. Palleschi, B. Lazzerini, *Spectrochim. Acta B: At. Spectrosc.*, **99**, 52 (2014).
35. J. P. Castro, D. V. Babos, E. R. Pereira-Filho, *Talanta*, **208**, 120443 (2020).
36. J. Zhang, R. Li, X. Zhang, Y. Bai, P. Cao, P. Hua, *Sci. Total Environ.*, **649**, 1314 (2019).
37. X. Jin, Y. Zhang, D. Yao, *Int. Symposium on Neural Networks*, 1022–1031 (2007).
38. X. Cui, Q. Wang, Y. Zhao, X. Qiao, G. Teng, *Appl. Phys. B*, **125**, 56 (2019).
39. S. Chutani, A. Goyal, *Multimed Tools Appl.*, **77**, 7447 (2018).
40. R. Noori, A. Karbassi, M. S. Sabahi, *J. Environ. Manage*, **91**, 767 (2010).
41. B. G. Alhassan, F. B. Yusof, S. M. Norrulashikin, *IJMH*, **4**, 40–48 (2020).
42. V. Singh, I. Gupta, H. Gupta, *Eng. Appl. Artific. Intell.*, **20**, 249 (2007).
43. S. K. Chaharsooghi, M. Rezaei, *Int. J. Serv. Oper. Manag.*, **25**, 120 (2016).
44. M. Dorofki, A. H. Elshafie, O. Jaafar, O. A. Karim, S. Mastura, *Int. Proc. Chem. Biol. Environ,* **33**, 39 (2012).
45. M. A. Shafi, N. Hamzah, *4th Int. Power Engineering and Optimization Conference*, 352 (2010).
46. F. Rezaei, *Plasma Science and Technology: Progress in Physical States and Chemical Reactions*, 363 (2016).
47. W. Zhang, S.-Y. Tang, Y.-F. Zhu, W.-P. Wang, *World Acad. Sci. Eng. Technol.*, **41**, 933 (2010).
48. D. Z. Antanasijević, M. D. Ristić, A. A. Perić-Grujić, V. V. Pocajt, *Int. J. Green. Gas Control*, **20**, 244 (2014).
49. I. Babaoğlu, O. Fındık, M. Bayrak, *Expert Syst. Appl.*, **37**, 2182 (2010).
50. V. M. Janakiraman, X. Nguyen, D. Assanis, *Appl. Soft Comput.*, **13**, 2375 (2013).