Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Nigde Omer Halisdemir University Journal of Engineering Sciences

ISSN: 2564-6605
Araştırma / Research

PREDICTION OF GROSS CALORIFIC VALUE OF COAL FROM PROXIMATE AND ULTIMATE ANALYSIS VARIABLES USING SUPPORT VECTOR MACHINES WITH FEATURE SELECTION

Mustafa AÇIKKAR[1,*]

[1] *Department of Aerospace Engineering, Faculty of Aeronautics and Astronautics, Adana Alparslan Türkeş Science and Technology University, Adana, Turkey*

ABSTRACT

The gross calorific value (GCV) is an essential thermal property of coal which indicates the amount of heat energy that could be released by burning a specific quantity. The primary objective of the presented study is to develop new GCV prediction models using support vector machines (SVMs) combined with feature selection algorithm. For this purpose, the feature selector RReliefF is applied to the dataset consisting of proximate and ultimate analysis variables to determine the importance of each predictor of GCV. In this way, seven different hybrid input sets (data models) were constructed. The prediction performance of models was computed by using the square of multiple correlation coefficient (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE). Considering all the results obtained from this study, the predictor variables moisture (M) and ash (A) obtained from the proximate analysis and carbon (C), hydrogen (H) and sulfur (S) obtained from the ultimate analysis were found to be the most relevant variables in predicting GCV of coal, while the predictor variables volatile matter (VM) from the proximate analysis and nitrogen (N) from the ultimate analysis did not have a positive effect on the prediction accuracy. The SVM-based model using the predictor variables M, A, C, H, and S yielded the highest R^2 and the lowest RMSE and MAPE with 0.998, 0.22 MJ/kg, and 0.66%, respectively. For comparison purposes, multilayer perceptron and radial basis function network were also used to predict GCV.

Keywords: Gross calorific value of coal, Feature selection, Support vector machines, Artificial neural network

ÖZELLİK SEÇİMİ İLE BİRLEŞTİRİLMİŞ DESTEK VEKTÖR MAKİNELERİNİ KULLANARAK KÖMÜRÜN ÜST ISIL DEĞERİNİN KISA VE ELEMENTEL ANALİZ DEĞİŞKENLERİNDEN TAHMİNİ

ÖZET

Üst ısıl değer (GCV), kömürün belirli bir miktarı yakıldığında açığa çıkan ısı enerjisi miktarını gösteren temel bir termal özelliğidir. Sunulan çalışmanın ana amacı, özellik seçimi algoritması ile destek vektör makineleri (SVM'ler) kullanarak yeni GCV tahmin modelleri geliştirmektir. Bu amaçla, literatürde ilk kez, özellik seçici RRelief-F algortiması, GCV'nin her bir tahmin edici değişkeninin önemini belirlemek için kısa ve elementel analiz değişkenlerinden oluşan veri kümesine uygulanmıştır. Bu şekilde, yedi farklı karma giriş seti (veri modelleri) oluşturulmuştur. Sunulan modellerin tahmin performansı, çoklu korelasyon katsayısının karesi (R^2), kök ortalama kare hatası (RMSE) ve ortalama mutlak yüzde hatası (MAPE) ile hesaplanmıştır. Bu çalışmadan elde edilen tüm sonuçlar değerlendirildiğinde, kısa analizden elde edilen nem (M) ve kül (A) ile elementel analizden elde edilen karbon (C), hidrojen (H) ve kükürt (S) değişkenleri kömürün GCV'sini tahmin etmede en uygun değişkenler olarak belirlenirken, kısa analizden elde edilen uçucu madde (VM) ile elementel analizden elde edilen nitrojenin (N) tahmin etme doğruluğu üzerinde olumlu bir etkiye sahip olmadığı görülmüştür. M, A, C, H ve S tahmin edici değişkenlerini kullanan SVM-tabanlı model, en yüksek R^2 ve en düşük RMSE ve MAPE değerlerini sırasıyla 0,998, 0,22 Mj/kg ve % 0,66 olarak vermiştir. Ayrıca, karşılaştırma amacıyla GCV'yi tahmin etmek için çok katmanlı algılayıcı ve radyal temelli fonksiyon ağı kullanılmıştır.

Anahtar kelimeler: Kömürün üst ısıl değeri, Özellik seçimi, Destek vektör makinesi, Yapay sinir ağları

*Sorumlu yazar / Corresponding author, e-posta / e-mail: macikkar@atu.edu.tr

Geliş / Recieved: 02.07.2019 Kabul / Accepted: 26.03.2020 doi: 10.28948/ngumuh.585596

1. INTRODUCTION

As a fossil fuel, coal is critical energy supplier material in many industries such as electricity generation, cement making, and coke generation for iron and steelmaking [1]. The reasons why coal is mostly used in energy generation to its abundance, and financial advantages and it is expected to remain the dominant energy source for the near future [2]–[4]. In particular, coal is projected to remain the mainstay of electricity generation in many major economies - especially where coal is the main indigenous and economically viable, source of energy [2]. Since coal is mostly used as an energy source, its heating capacity is important, and it is called as calorific value. The calorific value measured according to American Society for Testing and Materials (ASTM) standards is called gross calorific value (GCV) and sometimes higher heating value (HHV).

It can be said that the GCV is a function of coal compositions that can be obtained by proximate or ultimate analysis. Moisture (M), ash (A), volatile matter (VM), and fixed carbon (FC) are the variables of the proximate analysis. On the other hand, carbon (C), hydrogen (H), nitrogen (N), sulfur (S), and oxygen (O) are the coal compositions of the ultimate analysis [5]. The determination of GCV of coals with quick and accurate prediction models is one of the interesting studies in the literature. Engineers, designers, manufacturers, and researchers interested in the GCV of materials want to determine it in an easy and fast way within an acceptable tolerance. Therefore, a number of attempts have been made in the literature in order to develop a representative regression model [6]. Using several linear and nonlinear regression methods, researchers have developed models for GCV prediction based on proximate analysis results [1], [4], [7]–[11] and ultimate analysis results [5], [9], [10], [12], [13] as well as both analysis results [11], [14], [15].

The primary purpose of this study is to develop new GCV prediction models using proximate and ultimate analysis variables of coal. For this purpose, for the first time in the literature, the SVM and feature selection algorithm RRelief-F [16] have been used together to develop new GCV prediction models by identifying the relevant and irrelevant determinants of GCV. The main reason for using the feature selector is to determine and exclude unnecessary variables of the proximate and/or ultimate analysis from the dataset. Thus, the purpose is to increase the overall prediction accuracy when it is possible, to reduce the complexity of the developed prediction model as well as to decrease the cost of the calculation required to create the prediction model. The aforementioned feature selection algorithm was applied to the dataset to create several hybrid data models. In addition to these data models, two other data models that use only proximate and ultimate analysis variables were constructed separately for comparison purposes. To reveal the performance of the SVM on GCV prediction, two other artificial neural networks (ANNs) methods, namely, multilayer perceptron (MLP) and radial basis function network (RBFN) were also used to predict GCV on the dataset utilized. By calculating performance metrics square of multiple correlation coefficient (R^2), root mean square error (RMSE) and mean absolute percentage error (MAPE) of each developed model, the overall performance of SVM on GCV prediction were compared to that of MLP and RBFN.

2. DATASET GENERATION

The dataset used in this study was produced from the COALQUAL Version 3.0 database presented under U.S. Geological Survey Energy Resources Program [17]. The database consists of the analysis results of 6690 coal samples in as-received basis and includes proximate and ultimate analysis results as well as GCV of the samples. All analyses applied to the samples were made in accredited testing laboratories in accordance with ASTM standards. In order to provide consistency of the presented results, 15 samples where the validation status of proximate analysis and/or ultimate analysis was “Incomplete Data” were removed from the database.

In the proximate analysis, the values of M, A, and VM variables are obtained by measuring in the laboratory environment. On the other hand, the values of M, A, and VM variables are used to calculate FC variable by Equation 1. Similarly, in the ultimate analysis, while the values of C, H, N, and S variables are measured in the laboratory environment, the value of O variable is calculated by Equation 2.

$$FC = 100 - (M + A + VM) \tag{1}$$

$$O = 100 - (C + H + N + S + M + A) \tag{2}$$

In general, the values of the predictor variables used to form the dataset are expected to be obtained by measuring, not by calculating. If the value of a predictor variable depends only on the other predictor variables, the inclusion of that variable to the model does not produce a positive effect on the prediction performance of the model [18]. On the contrary, it increases the complexity of the model and the cost of the calculation required to create the prediction model.

As it is given in Equations 1 and 2, the value of FC directly depends on the values of the other proximate analysis variables, and the value of O depends on the other ultimate analysis variables, as well as the proximate analysis values of M and A, respectively.

Therefore, the proximate analysis variable FC and the ultimate analysis variable O are not included in the dataset. As a result, the dataset was generated using 6675 analysis results of the coal samples, and the proximate variables M, A and VM and the ultimate variables C, H, N, and S were determined as predictor variables to predict the target variable GCV.

The detailed information regarding the COALQUAL database can be found on the official website at http://ncrdspublic.er.usgs.gov/coalqual/.

The descriptive statistics of the generated dataset are given in Table 1. The scatter plots of target variable GCV according to the predictor variables are shown in Figure 1.

Table 1. Overview of the utilized dataset

Category	Type of Analysis	Variable Name	Minimum	Maximum	Mean	Standard Deviation
Predictor variables	Proximate	M (%)	0.40	52.50	8.15	9.99
		A (%)	0.90	54.70	11.79	7.29
		VM (%)	3.00	55.70	32.00	6.55
	Ultimate	C (%)	22.52	88.20	65.06	12.46
		H (%)	0.21	9.12	4.32	0.79
		N (%)	0.20	5.60	1.27	0.34
		S (%)	0.06	20.90	1.95	1.80
Target variable	-	GCV (Mj/kg)	8.81	36.25	26.89	5.36

3. METHODOLOGY AND PREDICTION MODELS

3.1. Methodology

In this study, the dataset composed of 6675 entries and described in detail in Section 2 was considered for experimental purposes. By conducting the RRelief-F algorithm for regression [16], the importance of each predictor variable in the utilized dataset was calculated separately and sorted in descending order according to the score of each predictor variable. Table 2 shows the ranking scores of each predictor variable calculated by the RRelief-F algorithm for the dataset used. Seven hybrid data models (Model 1 to 7) were created to predict GCV of coals by removing the variable, which had the lowest score from the list of predictor variables one by one and in order. In order to determine and clarify the effect of hybrid data models generated by the RRelief-F algorithm on the prediction accuracy, two other data models (Model 8 and 9), which were composed of proximate and ultimate analysis variables were constructed separately. Table 3 gives the data models and the predictor variables that each data model has.

Table 2. RRelief-F scores of the predictor variables for the utilized dataset.

Predictor Variable	RRelief-F Score
C	0.0105140
A	0.0038385
H	0.0037461
M	0.0029118
S	0.0020134
VM	0.0015223
N	0.0007616

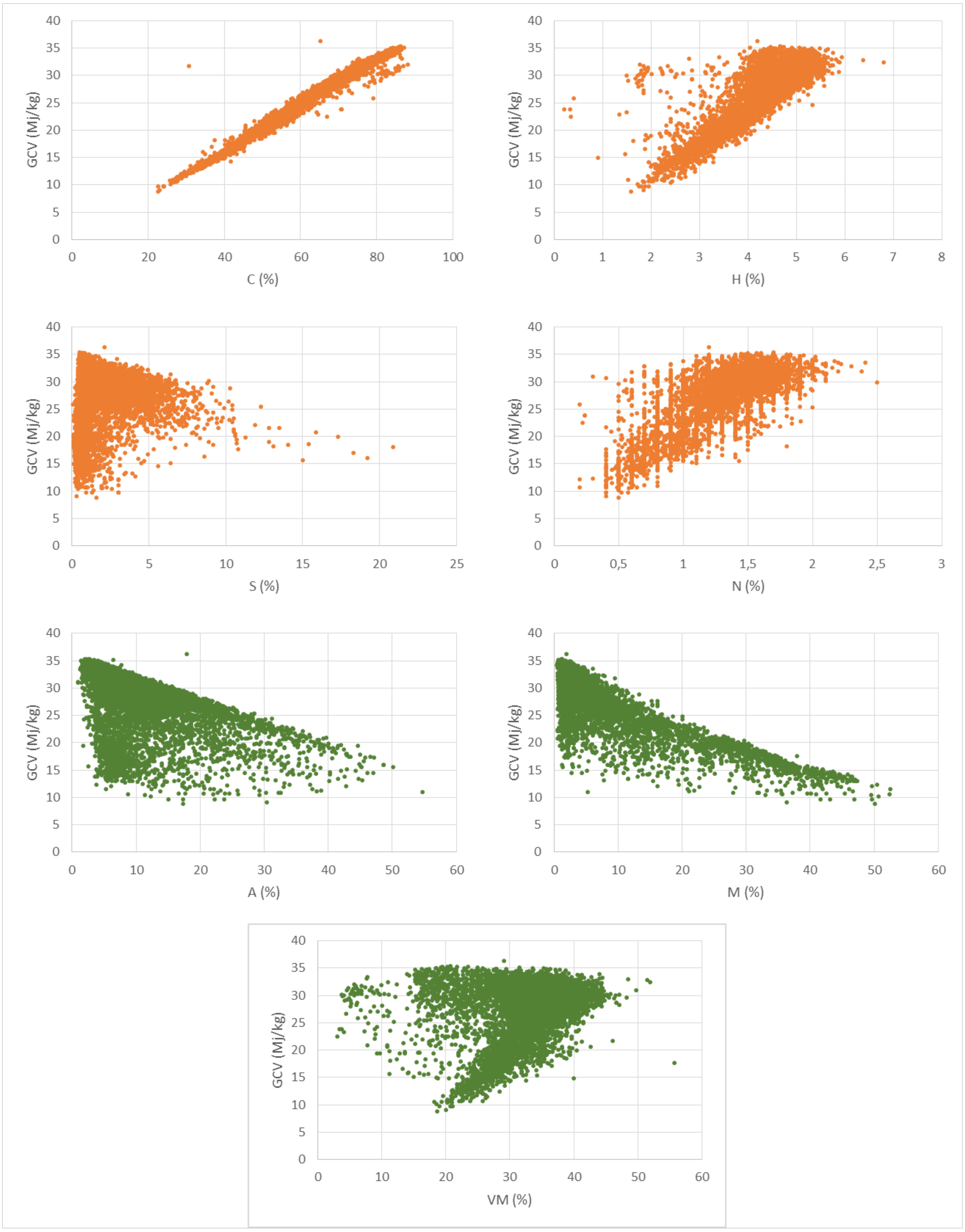

Figure 1. The scatter plots of GCV according to the predictor variables for the dataset.

Table 3. Overview of data models developed for GCV prediction

Type of Analysis	Model Name	Predictor Variables
Hybrid (obtained with RRelief-F algorithm)	Model 1	C, A, H, M, S, VM, N
	Model 2	C, A, H, M, S, VM
	Model 3	C, A, H, M, S
	Model 4	C, A, H, M
	Model 5	C, A, H
	Model 6	C, A
	Model 7	C
Ultimate	Model 8	C, H, S, N
Proximate	Model 9	A, M, VM

In this study, primarily, SVM-based prediction models were developed to predict the GCV of coal by using data models given in Table 3. In order to evaluate the prediction performance of the SVM-based models, the accuracy of these prediction models was compared with those of ANN-based prediction models. The choice of these machine learning methods for predicting GCV of coal is based on several factors. In the past, many studies [19]–[21] that showed the superiority of SVM over other regression methods were presented. Moreover, there are several studies using SVM with different datasets and data models [1], [4], [22]–[24]to predict the GCV of coals. On the other hand, ANNs have been widely used to solve real-world problems in different areas [18], [25], [26]. In this study, for comparison purposes, two types of ANNs, namely MLP and RBFN, were used to predict the GCV of coal. MLP is a well-known and most commonly used ANN method. In addition, RBFN is one of the other popular ANN methods and has been used successfully in the GCV prediction of coals from the proximate analysis results [18].

The generalization ability of the prediction models has been satisfied using 10-fold cross-validation, and the prediction performance of presented models was calculated by using R^2, RMSE, and MAPE. The formulas of these performance metrics are shown in Equations (3) - (5). It should be kept in mind that the lower RMSE and MAPE values and the higher R^2 value indicate that the prediction model is more accurate.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y_i'\right)^2} \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - Y_i'|}{Y_i} \tag{5}$$

In Equations (3) – (5), the target value is represented with Y, the predicted value is represented by Y', the mean of Y is represented by $\bar{Y}$ and the number of entries in a testing subset is given by n.

As stated above, in this study, 10-fold cross-validation was applied to each prediction model separately to strengthen the stability and reliability of the presented results. In k-fold cross-validation, the original data set is randomly partitioned into k equal size subsets. Of the k subsets, a single subset is retained as the testing data to validate the model, and the remaining k-1 subsets are used as training data. This procedure is repeated k times (the folds), with each of the k subsets used exactly once as the testing data. The advantage of this method is that all instances are used for both training and testing, and each instance is used for testing exactly once. The reported value of the given performance metrics for each prediction model was computed by taking the average of the values of the performance metrics obtained from each fold.

All the experiments in this study, including manipulating, analyzing, and modeling the data presented were done using DTREG [27], a predictive modeling software that implements the most widely used types of machine learning methods.

3.2. SVM-based prediction models

The performance of the SVM model for a given regression problem is directly related to the values of the model parameters. The value of epsilon (ε), the value of capacity (C), the type of kernel function, and the parameter of the kernel function, if any, should be determined carefully. Choosing the type of kernel function and determining the optimal parameter values of this function are very important for creating SVM-based models [28]. Kernel functions commonly used in SVM are generally divided into four categories as linear, polynomial, radial basis function (RBF) and sigmoid kernels. In the literature, the RBF kernel has mostly been used to present satisfactory generalization capabilities [21], [29], [30]. For this reason, in this study, RBF was used as the kernel function. In the RBF kernel function, the parameter γ must also be optimized. In order to develop SVM-based models with satisfactory prediction accuracy, the most appropriate triple of (C, ε, γ) needs to be revealed for the given regression problem.

The prediction accuracy of SVM is very sensitive to the selection of the parameters, and there is no rule or formula based on a mathematical model for obtaining the exact desired values of those parameters. For this purpose, a suitable parameter optimization algorithm should be used to find the optimal values of the parameters used for SVM-based models, so that good generalization performance could be achieved. Therefore, one needs an effective search algorithm to find the best values of these parameters. The grid search (GS) [31] is one of the most commonly used algorithms for optimizing the parameters of SVM. The value of each parameter is changed within a pre-defined range. The range of each parameter is divided by the number of steps on a logarithmic scale, and the GS tries each set of hyper-parameters severally. At the end of the search process, the parameters that provide the highest accuracy are selected as optimized parameters.

The range of the hyper-parameters, the number of steps and the scale type of the hyper-parameters utilized in each SVM-based model are shown in Table 4.

3.3. ANN-based prediction models

An artificial neural network (ANN) is a flexible mathematical structure, which is capable of identifying complex nonlinear relationships between input and output data [32]. ANNs have been widely applied to solve many difficult problems in different areas. There are many types of ANNs for modeling function approximation of the problems. Two types of neural networks, MLP and RBFN, were used to verify the utility of the proposed prediction models in this study.

3.3.1. MLP-based prediction models

MLP, which is composed of one input layer, one or more hidden layers, and one output layer, is a flexible, general-purpose, and most widely used network within ANN methods. Finding the optimal number of neurons in each hidden layer is one of the most important issue of MLP. If an insufficient number of neurons or more than necessary neurons are used, the created network will be unable to model data, and its performance will be poor. There is no theoretical approach to determine the optimal number of neurons in the hidden layer, which varies with the type and complexity of the problem.

For this study, one hidden layer was used to construct the MLP network. The logistic transfer function was set as the activation function in the hidden layer as well as the output layer. The training method of the MLP network is the scaled conjugate gradient algorithm. The optimal number of neurons in the hidden layer was found by comparing the prediction performance of the networks generated using each value in the range of neurons given in Table 4. This approach, which is computationally expensive due to having to build many models, is a highly effective method for finding the optimal number of neurons.

3.3.2. RBFN-based prediction models

RBFN is a special case of multilayer feed-forward neural network consisting of three layers including one input layer, one hidden layer, and one output layer; and is different in terms of node characteristics and learning algorithm [33]. The neurons in the hidden layer contain RBF transfer functions whose outputs are inversely proportional to the distance from the center of the neuron [27].

Four different parameters are determined in the training phase of the RBFN model. These parameters are the number of neurons in the hidden layer, the center points of each hidden-layer RBF function, the radius (spread) of each RBF function in each dimension, and the weights applied to the RBF function outputs as they are passed to the output layer [27].

For this study, the optimal number of neurons in the hidden layer was found by comparing the prediction performance of the networks generated using each value in the range of neurons given in Table 4. The algorithm developed by Sheng Chen et al. [34] was used as the training method in the RBFN model. This method determines the optimal center points and spreads for each neuron with an evolutionary approach. Ridge regression is utilized for the computation of the optimal weights between the neurons in the hidden layer and the output layer. The optimal regularization Lambda parameter that minimizes the generalized prediction error is computed by an iterative procedure developed by M. J. L. Orr [35] [27].

The flowchart of the regression-based prediction model using SVM, MLP, and RBFN is illustrated in Figure 2.

Table 4. Values of the utilized parameters for the regression methods

Method	Parameter	Value
SVM	Capacity (*C*)	[0.1, 10000]
	Epsilon (ε)	[0.0001, 100]
	Gamma (γ)	[0.001, 50]
	Number of steps for GS	15
	Scale type of the hyper-parameters for GS	Logarithmic-scaled
MLP	Number of neurons in the hidden layer	[2, 15]
	Hidden layer/output layer transfer function	Logistic
	Training method	Scaled conjugate gradient
RBFN	Maximum number of neurons in the hidden layer	200
	The radius (spread) of each RBF function in each dimension	[0.01, 100]
	Regularization parameter (Lambda)	[0.001, 10]
	Training method	Orthogonal forward selection [34]

4. RESULTS AND DISCUSSION

4.1. Results

To determine the effectiveness of the presented GCV prediction models based on the hybrid, proximate and ultimate analysis variables, experiments were conducted on the data models (Model 1 to 9) using SVM, MLP, and RBFN methods. Tables 5 – 7 show the R^2, RMSE, and MAPE values obtained from GCV prediction models generated by each data model and regression method.

Table 5. Performance measures of SVM-based models for GCV prediction.

Data Model	Predictor Variables	Training Set			Testing Set		
		R^2	RMSE (Mj/kg)	MAPE (%)	R^2	RMSE (Mj/kg)	MAPE (%)
Model 1	C, A, H, M, S, VM, N	0.998	0.21	0.62	0.998	0.22	0.65
Model 2	C, A, H, M, S, VM	0.998	0.22	0.62	0.998	0.22	0.66
Model 3	C, A, H, M, S	0.998	0.22	0.63	0.998	0.22	0.66
Model 4	C, A, H, M	0.996	0.30	0.86	0.996	0.31	0.90
Model 5	C, A, H	0.996	0.35	1.00	0.995	0.36	1.03
Model 6	C, A	0.992	0.46	1.29	0.992	0.47	1.31
Model 7	C	0.986	0.62	1.68	0.989	0.56	1.68
Model 8	C, H, S, N	0.998	0.25	0.72	0.998	0.26	0.75
Model 9	A, M, VM	0.986	0.60	1.74	0.986	0.61	1.78

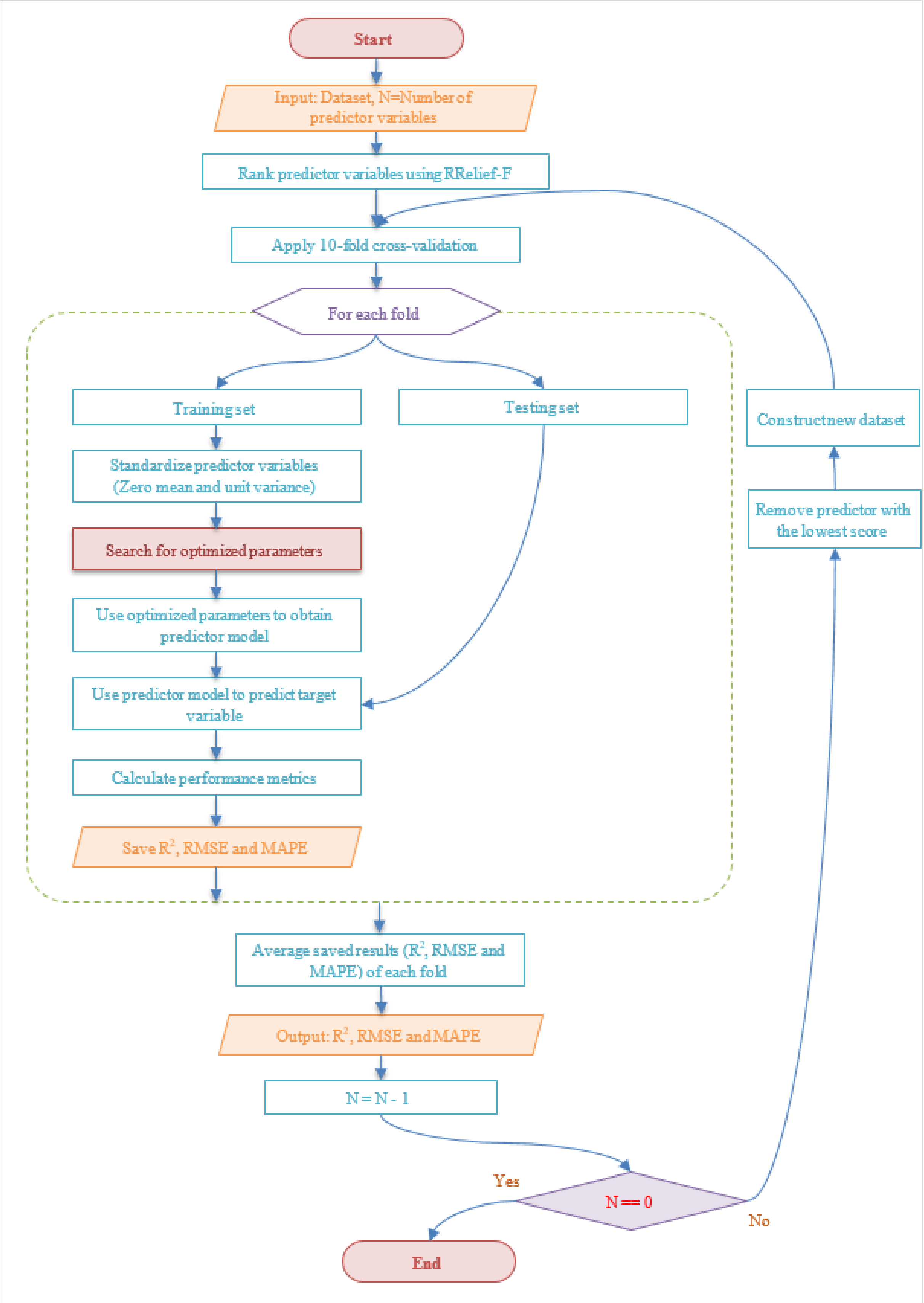
Start
Input: Dataset, N=Number of predictor variables
Rank predictor variables using RRelief-F
Apply 10-fold cross-validation
For each fold
Training set
Testing set
Standardize predictor variables (Zero mean and unit variance)
Search for optimized parameters
Use optimized parameters to obtain predictor model
Use predictor model to predict target variable
Calculate performance metrics
Save R², RMSE and MAPE
Average saved results (R², RMSE and MAPE) of each fold
Output: R², RMSE and MAPE
N = N - 1
N == 0
Yes
No
End
Construct new dataset
Remove predictor with the lowest score

Figure 2. Flowchart of the models based on the hybrid data models generated by the RRelief-F algorithm

Table 6. Performance measures of MLP-based models for GCV prediction.

Data Model	**Predictor Variables**	**Training Set**			**Testing Set**		
		R^2	**RMSE (Mj/kg)**	**MAPE (%)**	**R^2**	**RMSE (Mj/kg)**	**MAPE (%)**
Model 1	C, A, H, M, S, VM, N	0.997	0.26	0.76	0.997	0.26	0.75
Model 2	C, A, H, M, S, VM	0.997	0.25	0.75	0.997	0.26	0.77
Model 3	C, A, H, M, S	0.997	0.26	0.75	0.997	0.26	0.76
Model 4	C, A, H, M	0.995	0.34	1.03	0.995	0.35	1.08
Model 5	C, A, H	0.995	0.36	1.08	0.995	0.37	1.12
Model 6	C, A	0.990	0.52	1.46	0.990	0.53	1.46
Model 7	C	0.986	0.61	1.68	0.986	0.63	1.70
Model 8	C, H, S, N	0.997	0.26	0.78	0.997	0.27	0.81
Model 9	A, M, VM	0.986	0.64	1.87	0.985	0.65	1.92

Table 7. Performance measures of RBFN-based models for GCV prediction.

Data Model	**Predictor Variables**	**Training Set**			**Testing Set**		
		R^2	**RMSE (Mj/kg)**	**MAPE (%)**	**R^2**	**RMSE (Mj/kg)**	**MAPE (%)**
Model 1	C, A, H, M, S, VM, N	0.998	0.22	0.66	0.998	0.24	0.68
Model 2	C, A, H, M, S, VM	0.998	0.22	0.65	0.998	0.23	0.68
Model 3	C, A, H, M, S	0.998	0.23	0.66	0.998	0.23	0.68
Model 4	C, A, H, M	0.996	0.30	0.91	0.996	0.32	0.94
Model 5	C, A, H	0.995	0.34	1.02	0.995	0.36	1.04
Model 6	C, A	0.992	0.46	1.32	0.992	0.47	1.34
Model 7	C	0.986	0.61	1.68	0.986	0.62	1.69
Model 8	C, H, S, N	0.998	0.25	0.73	0.997	0.26	0.76
Model 9	A, M, VM	0.987	0.60	1.73	0.986	0.62	1.78

4.2. Discussion

The results concerning the training set shown in Tables 5 – 7 were given for information purposes, and all comparisons were made with respect to the results of the testing set. According to the results, the following comments can be asserted for the prediction of the GCV and the relevant and irrelevant predictors affecting the accuracy of prediction models:

- The results presented in this study reveal that SVM-based models outperform MLP- and RBFN-based models for GCV prediction. In more detail, the average RMSE's of all SVM-based prediction models are 3.71% and 10.83% lower than the ones of RBFN-based and MLP-based models, respectively. In addition, in ANN-based prediction models, RBFN-based prediction models give lower RMSE's than the MLP-based prediction models for GCV prediction. It should be noted that the performance of the SVM-based and RBFN-based models is almost similar to each other for all data models.
- The results obtained from the prediction models using data models generated by the RRelief-F algorithm (Model 1 to 7) show that the ultimate analysis variable C and N are the most relevant and irrelevant predictors of GCV, irrespective of the regression method used, respectively.
- When the effect of the predictor variables is examined on prediction performance, the inclusion of the predictor variables VM and N yields no effect or a negligible effect on prediction, regardless of the regression method utilized.
- Among the Model 1 to 7, Model 3, which contains the proximate analysis variables M and A; the ultimate analysis variables C, H, and S gives the lowest RMSE's with 0.22 Mj/kg, 0.23 Mj/kg and 0.26 Mj/kg, whereas Model 7 including only the ultimate analysis variable C yields the highest RMSE's with 0.56 Mj/kg, 0.62 Mj/kg and 0.63 Mj/kg for SVM, RBFN, and MLP, respectively.
- The RMSE's of the prediction models using the best hybrid data model (Model 3) are 18.18%, 13.04%, and 3.84% lower than the RMSE's of the prediction models using the ultimate analysis variables (Model 8) for SVM, RBFN, and MLP, respectively.

- The RMSE's of the prediction models using the best hybrid data model (Model 3) are 177.27%, 169.56%, and 150.00% lower than the RMSE's of the prediction models using the proximate analysis variables (Model 9) for SVM, RBFN, and MLP, respectively.

4.3. Comparing the results

As mentioned previously, the literature includes many studies using linear and nonlinear regression methods for predicting GCV of coals based on proximate and ultimate analysis results as well as both analysis results. When an evaluation is made considering the scope of these studies, each study differs from each other methodologically, and this difference can be grouped under two main headings. Firstly, the dataset utilized by each study may contain different number of samples, may include the results of the analysis of coal extracted from different regions and of different quality, and may have different predictor variables depending on the basis and/or the type of the analysis. Secondly, each study may use different regression methods, different hyper-parameter optimization algorithms concerning each regression method if operable, different hyper-parameters if any, different validation methods with various data splitting rules, and different metrics to determine prediction accuracy. Therefore, it is not applicable and reasonable to compare the results of the present study one-to-one, directly and in detail with the result of these studies in the literature, especially if the dataset used is different. However, without making a direct comparison with other studies, it can be inferred that the results of this study are relatively valuable by evaluating the results obtained from other studies.

The studies [9, 11] have shown that the predictive models using the ultimate analysis variables yield better results than the prediction models using the proximate analysis variables for the same dataset, regardless of which regression method is used. Similarly, in the present study, when the performance of the models using only the ultimate and proximate analysis variables was investigated, it was found that the models using the ultimate analysis variables (Model 8) were more accurate than the models using the proximate analysis variables (Model 9). At the same time, the present study also proved that the RMSE value of the SVM-based model using hybrid data obtained with the RRelief-F algorithm (Model 3) was 18.18% lower than that of the developed model using the ultimate analysis variables.

Hadavandi et al. [23] used a dataset with different coal analyses (proximate, ultimate, various sulfur forms, and petrography) of 924 coal samples, and developed a new SVM-based prediction model by applying Variable Importance Measurement (VIM). In VIM algorithm, the details of which is given in the study [23], the Relative Variable Importance (RVI) score of each input variable is calculated by creating SVM-based models. For comparison purposes, the proposed VIM algorithm by Hadavandi et al. [23] implemented to the utilized dataset in this study. Firstly, the RVI score of each predictor variable was calculated using SVM separately and ranked in descending order in compliance with their scores to find which ones are most important for the modeling. Secondly, seven VIM-based data models (VIM Model 1 to 7) were created by removing the variable, which had the lowest score one by one and in order. Finally, SVM-based prediction models were developed to predict the GCV of coal by using VIM data models. Table 8 gives the RVI scores of the predictor variables. Table 9 shows the VIM data models and the predictor variables that each model has as well as the R^2, RMSE, and MAPE values of SVM-based GCV prediction models.

Table 8. RVI scores of the predictor variables for the utilized dataset.

Predictor Variable	RVI Score
C	0,71906232
S	0,14892001
H	0,06820786
M	0,02700789
A	0,01961223
N	0,01193847
VM	0,00525122

Table 9. Performance measures of SVM-based models obtained by VIM for GCV prediction.

Data Model	Predictor Variables	Training Set			Testing Set		
		R^2	RMSE (Mj/kg)	MAPE (%)	R^2	RMSE (Mj/kg)	MAPE (%)
VIM Model 1	C, S, H, M, A, N, VM	0.998	0.21	0.62	0.998	0.22	0.65
VIM Model 2	C, S, H, M, A, N	0.998	0.23	0.64	0.998	0.23	0.67
VIM Model 3	C, S, H, M, A	0.998	0.22	0.63	0.998	0.22	0.66
VIM Model 4	C, S, H, M	0.997	0.27	0.82	0.997	0.28	0.83

VIM Model 5	C, S, H	0.996	0.28	0.83	0.996	0.29	0.85
VIM Model 6	C, S	0.992	0.48	1.20	0.993	0.49	1.20
VIM Model 7	C	0.986	0.62	1.68	0.989	0.56	1.68

The following comments and comparisons can be made for the results regarding SVM-based prediction models obtained by RRelief-F and VIM algorithms:

- In VIM algorithm, it is necessary to create total of (*m*+1) SVM-based models in order to calculate the RVI scores of the predictor variables where *m* is the number of predictor variables. Therefore, the execution time of the VIM algorithm strongly depends on the number of predictor variables, and the size and complexity of the dataset. As a result, the VIM algorithm consumes extremely much more time since it has to develop (*m*+1) SVM-based models, while the RRelief-F algorithm produces results in less than a second.
- RRelief-F algorithm and VIM algorithm ranked the variables in a different order as seen in Table 2 and Table 8, respectively. Although the variables forming Model 3 and VIM Model 3 are in a different order, they have the same predictor variables (C, A, H, M, S). In addition, both algorithms showed similarly that C is the most important variable.
- When the results of SVM based models given in Table 5 and Table 9 were examined, Model 3 and VIM Model 3, which yield the lowest RMSE (0.22 Mj/kg) with the least number of predictor variables, were the best models.
- Considering that Model 3 and VIM Model 3 are the best models, it was seen that N and VM predictive variables had no effect on the accuracy of the prediction models.
- Considering only N and VM, the results of Model 2 and VIM Model 2 (0.22 Mj/kg and 0.23 Mj/kg) showed that the ranking made by the RRelief-F algorithm was more accurate and acceptable than that of the VIM algorithm.

As a result, taking into account the results and findings above, it can be concluded that the new hybrid SVM-based model provides promising results in predicting GCV of coal. Additionally, when the predictor variables proposed in this study based on proximate and ultimate analysis are used, it is considered that the accuracy of the prediction model will increase relatively regardless of which dataset will be used.

5. CONCLUSIONS

The aim of this study was to develop new GCV prediction models and bring out the distinctive predictor variables of GCV using SVM together with the feature selection algorithm. The dataset was formed by combining the proximate and ultimate analyses results of the 6675 coal samples in as-received basis. By applying feature selection algorithm RRelief-F to the dataset, seven hybrid data models were created for GCV prediction. In addition to these data models, two data models utilizing proximate and ultimate analysis results were also created for comparison purposes. Besides, two different ANN methods named MLP and RBFN were applied to the data models in order the make comparison with SVM. 10-fold cross-validation was used to satisfy the generalization capability of the developed models. The performances of the SVM-, MLP-, and RBFN-based prediction models were obtained by calculating the values of R^2, RMSE, and MAPE metrics.

Among the presented regression models, SVM-based prediction models performed better than the models developed by using MLP or RBFN irrespective of which data model was used. However, it should be taken into account that the prediction performance of the RBFN-based prediction models is almost as good as that of SVM-based prediction models. In addition, regardless of the regression method used, the same comments can be made for the effect of predictor variables for all the results related to the presented regression methods.

Considering the performances of all the models presented in this study, the results reveal that, VM and N have no positive effect on the performance of GCV prediction. Besides, the SVM-based predictor model that includes the predictor variables M and A obtained from the proximate analysis and C, H, and S obtained from the ultimate analysis is the most appropriate model for predicting GCV of coal in as-received basis.

REFERENCES

[1] Q. Feng, J. Zhang, X. Zhang, and S. Wen, “Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating conditional expectation and artificial neural network,” *Fuel Processing and Technology*, vol. 129, pp. 120–129, Jan. 2015.

[2] A. Garg and P. R. Shukla, “Coal and energy security for India: Role of carbon dioxide (CO_2) capture and storage (CCS),” *Energy*, vol. 34, no. 8, pp. 1032–1041, 2009.

[3] W. Chen and R. Xu, “Clean coal technology development in China,” *Energy Policy*, vol. 38, no. 5, pp. 2123–2130,

2010.

[4] P. Tan, C. Zhang, J. Xia, Q. Y. Fang, and G. Chen, "Estimation of higher heating value of coal based on proximate analysis using support vector regression," *Fuel Processing and Technology*, vol. 138, Oct., pp. 298–304, 2015.

[5] S. U. Patel *et al.*, "Estimation of gross calorific value of coals using artificial neural networks," *Fuel*, vol. 86, no. 3, pp. 334–344, 2007.

[6] A. V. Akkaya, "Proximate analysis based multiple regression models for higher heating value estimation of low rank coals," *Fuel Processing and Technology*, vol. 90, no. 2, pp. 165–170, 2009.

[7] A. K. Majumder, R. Jain, P. Banerjee, and J. P. Barnwal, "Development of a new proximate analysis based correlation to predict calorific value of coal," *Fuel*, vol. 87, no. 13–14, pp. 3077–3081, 2008.

[8] S. Yerel and T. Ersen, "Prediction of the Calorific Value of Coal Deposit Using Linear Regression Analysis," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 35, no. 10, pp. 976–980, May 2013.

[9] S. S. Matin and S. C. Chelgani, "Estimation of coal gross calorific value based on various analyses by random forest method," *Fuel*, vol. 177, pp. 274–278, 2016.

[10] S. Mesroghli, E. Jorjani, and S. C. Chelgani, "Estimation of gross calorific value based on coal analysis using regression and artificial neural networks," *International Journal of Coal Geology*, vol. 79, no. 1–2, pp. 49–54, 2009.

[11] X. Wen, S. Jian, and J. Wang, "Prediction models of calorific value of coal based on wavelet neural networks," *Fuel*, vol. 199, pp. 512–522, 2017.

[12] I. Yilmaz, N. Y. Erik, and O. Kaynar, "Different types of learning algorithms of artificial neural network (ANN) models for prediction of gross calorific value (GCV) of coals," *Scientific Research and Essays*, vol. 5, no. 16, pp. 2242–2249, 2010.

[13] I. Boumanchar *et al.*, "Multiple regression and genetic programming for coal higher heating value estimation," *International Journal of Green Energy*, vol. 15, no. 14–15, pp. 958–964, 2018.

[14] A. K. Verma, T. N. Singh, and M. Monjezi, "Intelligent prediction of heating value of coal," *Iranian Journal of Earth Sciences*, vol. 2, pp. 32–38, 2010.

[15] S. C. Chelgani, S. Mesroghli, and J. C. Hower, "Simultaneous prediction of coal rank parameters based on ultimate analysis using regression and artificial neural network," *International Journal of Coal Geology*, vol. 83, no. 1, pp. 31–34, 2010.

[16] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1/2, pp. 23–69, 2003.

[17] C. Palmer, C. Oman, A. Park, and J. Luppens, "The U.S. Geological Survey Coal Quality (COALQUAL) Database Version 3.0 Data Series 975," p. 43, 2015.

[18] M. Acikkar and O. Sivrikaya, "Prediction of gross calorific value of coal based on proximate analysis using multiple linear regression and artificial neural networks," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 5, pp. 2541–2552, 2018.

[19] V. H. Quej, J. Almorox, J. A. Arnaldo, and L. Saito, "ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 155, pp. 62–70, Mar. 2017.

[20] K. O. Akande, T. O. Owolabi, S. Twaha, and S. O. Olatunji, "Performance Comparison of SVM and ANN in Predicting Compressive Strength of Concrete," *IOSR Journal of Computer Engineering*, vol. 16, no. 5, pp. 88-94, 2014.

[21] F. Abut, M. F. Akay, and J. George, "Developing new VO2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection," *Computers in Biology and Medicine*, vol. 79, no. October, pp. 182–192, 2016.

[22] J. Fu, "Application of SVM in the estimation of GCV of coal and a comparison study of the accuracy and robustness of SVM," in *2016 International Conference on Management Science and Engineering (ICMSE)*, 2016, pp. 553–560.

[23] E. Hadavandi, J. C. Hower, and S. C. Chelgani, "Modeling of gross calorific value based on coal properties by support vector regression method," *Modeling Earth Systems and Environment*, vol. 3, no. 1, p. 37, Apr. 2017.

[24] M. Qi, H. Luo, P. Wei, and Z. Fu, "Estimation of low calorific value of blended coals based on support vector regression and sensitivity analysis in coal-fired power plants," *Fuel*, vol. 236, pp. 1400–1407, 2019.

[25] X. Huang, X. Liu, and Y. Ren, "Enterprise credit risk evaluation based on neural network algorithm," *Cognitive Systems Research*, vol. 52, pp. 317–324, Dec. 2018.

[26] S. Ren and L. Gao, "Combining artificial neural networks with data fusion to analyze overlapping spectra of nitroaniline isomers," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 2, pp. 276–282, Jul. 2011

[27] P. H. Sherrod, "DTREG Predictive Modeling Software," 2003.

[28] Ö. Baydaroğlu and K. Koçak, "SVR-based prediction of evaporation combined with chaotic approach," *Journal of Hydrology*, vol. 508, pp. 356–363, Jan. 2014.

[29] C. Campbell, "Kernel methods: a survey of current techniques", 2002.

[30] T. Kavzoglu and I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification,"

International Journal of Applied Earth Observation and Geoinformation, vol. 11, no. 5, pp. 352–359, Oct. 2009.

[31] C. W. Hsu, C. C. Chang, and C.J. Lin, “A Practical Guide to Support Vector Classification”, 2003.

[32] G. Ozbayoglu, A. M. Ozbayoglu, and M. E. Ozbayoglu, “Estimation of Hardgrove grindability index of Turkish coals by neural networks,” *International Journal of Mineral Processing*, vol. 85, pp. 93-100, 2003.

[33] J. Arliansyah and Y. Hartono, “Trip Attraction Model Using Radial Basis Function Neural Networks,” *Procedia Engineering*, vol. 125, pp. 445–451, Jan. 2015.

[34] S. Chen, X. Hong, and C.J. Harris, “Orthogonal Forward Selection for Constructing the Radial Basis Function Network with Tunable Nodes,” In *International Conference on Intelligent Computing*, Springer, Berlin, Heidelberg, August 23-26, 2005, pp. 777–786,.

[35] M. J. L. Orr, “Introduction to Radial Basis Function Networks,” Centre for Cognitive Science, University of Edinburgh, Scotland, 1996.