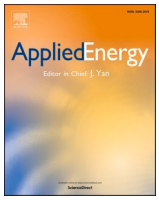
Applied Energy 347 (2023) 121499

Contents lists available at ScienceDirect

Applied Energy

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



A data-driven approach for the prediction of coal seam gas content using machine learning techniques

Satuk Buğra Akdaş [a,*], Abdullah Fişne [a]

[a] İTÜ Maden Fakültesi Maden Mühendisliği, Bölümü 34469 Maslak, İstanbul Sarıyer, İstanbul, Turkey

HIGHLIGHTS

- Novel data-driven solutions are given for predicting coal gas content.
- Models were constructed via Machine Learning methods.
- The model is applied for a low rank coal samples and account with various coal types.
- The Support Vector Machine is determined as successful tool for predicting of gas content with coal system variables.
- A sensitivity study using the model shows the effect of the coal features on gas content prediction.

ARTICLE INFO

Keywords:
Machine learning
Coal gas content
Support vector machine
Artificial neural network

ABSTRACT

A new data-driven approach to interpreting the nonlinear problem of total desorbed gas content analysis of coal seams is presented in this study. The study focuses on a low-rank coal reserve located in the Kınık coalfield, which was investigated using the United States Bureau of Mines (USBM) direct desorption method to anticipate the total desorbed gas content of coal seams for underground mining operations. The core samples collected during the reserve and gas content analysis were used to feed machine learning models, which were trained using coal properties data such as depth, moisture, ash, volatile matter, and calorific value, in relation to total desorbed gas content. Multiple linear regression, support vector machine, and artificial neural network were employed to predict the total desorbed gas content of coal seams in the Kınık coalfield. The machine learning models were optimized using hyperparameter tuning, and the most successful model was selected based on its regression and computational cost performance. Sensitivity analysis was conducted to investigate the performance of the coal properties on total desorbed gas content. The selected model was then utilized for predicting the total desorbed gas content of coal seams at a single point in the coalfield. The findings of this study provide insights and guidelines for unconventional reservoir analysis and petrophysical system prediction using machine learning methods. Overall, this study demonstrates the potential of machine learning in addressing nonlinear problems in the field of geology and provides a promising approach for future research in this area.

1. Introduction

The roots of the modern industry can be driven back to the use of coal-powered machines during the Industrial Revolution. However, with the continuous changes in the quantity and quality of energy sources, the focus has shifted towards more efficient and cleaner alternatives for a sustainable future. Although coal remains a major source of energy for heating, household, transportation, and industrial uses, its significant greenhouse gas emissions have prompted a move toward natural gas as a replacement. However, traditional utilization of coal sources and reclamation of by-products from coal mining continues to be a serious problem. According to BP's world energy reviews, as of 2020, there were proven coal reserves of 1.074×10^{12} tonnes worldwide [1], with some countries having more than 500 years of reserve use relative to production. The Kyoto Protocol, which was implemented in 2009, has placed restrictions on the use of coal due to concerns about global warming and low carbon emission targets. Although natural gas, renewable energy sources, and nuclear power can supply the fundamental needs of energy production, heavy industries such as steel and glass production remain dependent on high and continuous calorific

* Corresponding author.
E-mail addresses: akdass@itu.edu.tr (S.B. Akdaş), fisnea@itu.edu.tr (A. Fişne).

https://doi.org/10.1016/j.apenergy.2023.121499
Received 16 June 2022; Received in revised form 6 June 2023; Accepted 19 June 2023
Available online 25 June 2023

value sources.

In late 2019, the world was hit by the Covid-19 pandemic, which resulted in reduced mobility and productivity starting from May 2020. This interruption affected oil and gas reservoir management, mining activities, and energy raw material transportation, leading to chaos in energy production phases. Supply chain disruptions along with the variety of suppliers caused the prices of oil and gas arising to over $100 per barrel. Furthermore, political, and regional conflicts on mainstream natural gas producers and distribution networks have intensified the negative impact. The surge in the energy market has also significantly impacted coal prices. In April 2022, coal prices peaked at $439 [2] due to the shortage of natural gas supply. Despite its greenhouse gas emission concerns, coal emerged as an unavoidable source in the absence of sufficient natural gas supply.

Coalbed methane is a key energy source in the transition from coal to natural gas and eventually to renewable energy sources [3 4]. The gas content of coal is a critical factor in this transition. Similar to unconventional oil and gas reservoirs [5], coal has unique properties that make it an attractive energy source [6]. Coal not only provides heat when burned, but it also contains a mixture of gas components, predominantly methane, which are formed during its coalification process. Over millions of years, the buried coal seams become enriched with gas content, which can be mostly captured. While a significant amount of gas may escape during the coalification process, enough gas can be bound within coal's porous media.

Exploration and production of coalbed methane (CBM) began in North America during the 1980s [7]. The total estimated reserves of coalbed gas in the United States are around 19.8×10^{12} m^3, of which 14.2% are considered recoverable. The US Department of Energy has published proven reserves of coalbed methane as 336.34 Bcm [8]. Similarly, countries such as China [9] and Australia [10] have also shown potential for coalbed methane prospects. However, the CBM potential of coal reserves is dynamic and constantly changing, which necessitates the evaluation of coal gas content as an alternative natural gas source. Therefore, it is essential to determine the amount of gas content in each section of the underground system that may be encountered for effective operational design due to the importance of coal gas.

Gas content of coal seam has been defined as the volume of gas sorbed within the micropores in unit mass of coal [11 12]. Gas content determination methods can be grouped into two categories [12,13]. The first is direct methods, which measure the volume of gas released from a coal sample in a desorption canister. Second one is indirect methods, which are based on gas sorption characteristics under given temperature and pressure conditions. Those processes were evaluated in Table 1.

The direct gas content determination methods subdivide the total gas content of a coal sample into three components [12]. These components are defined as lost (Q_1), desorbed (Q_2) and residual (Q_3) gas. The Q_1 is gas lost from the samples subsequent to its removal from its in-situ position and prior to its containment in the canister. Q_2 is the amount of gas desorbed in the container after sealing the sample, transportation to the laboratory and before pulverizing the sample. Q_3 is the gas still contained in the coal sample before its pulverization. Each of these components are generally measured or estimated by a different procedure, and then combined to yield the total gas content of the sample [14].

Table 1 summarizes the measurement methods and limitations, which are all characterized by delicate sampling, long, and expensive experimental processes. In this study, we aim to develop a guide for predicting the total desorbed gas content based on a series of data previously collected from Kınık coalfield using the USBM direct method [12]. To achieve a faster and less expensive gas content determination, we propose the use of computational power and machine learning methods. In the following section, we describe the geological system of Kınık coalfield, which is complex and sensitive to geological and artificial interferences such as faults, formations, coal seams, exploited open pit mine sections, burial depths, etc. Given the complexity, it is challenging to adopt a homogeneous approach for coal gas prediction. However, low permeable structures and features present in coal samples can help in building a data-driven model.

Table 1
Desorbed Gas Measurement Techniques retrieved from Hou et al. [15].

Method	Name	Applied Conditions	Limitations
Direct method	USBM direct method	Known desorption and residual gas data	Long desorption time
		Known reservoir temperature	Estimated lost gas time result in inaccurate lost gas volume
			Integrated cores are required
	Curve Fit method	Known desorption data	Long desorption time
		Known reservoir temperature	Estimated lost gas and residual gas volume
			Require integrated cores
			Ignoring desorption difference between desorbed and residual gases
			Residual gas is assumed to be negligible
	Smith and Williams method	Known desorption data	Gas content is estimated
		Known reservoir temperature	Integrated cores are required
			Ignoring desorption difference between desorbed and residual gases
			Residual gas is assumed to be negligible
Indirect method	Sorption isotherm	Variable gas pressure and certain temperature	Strong assumption of gas saturation
		Known both reservoir pressure and temperature	Limitation caused by temperature

Reserve determination and core sampling are direct methods that involve interacting with the underground system. In coal reserves, it is crucial to obtain comprehensive information about the geological aspects, estimate reserve quantities, and assess coal quality through core drilling operations. However, when conducting gas content research, additional coal core samples are necessary to capture possible entrapped gas. These additional core samples incur additional costs during operations. Moreover, gas content determination does not conclude with the completion of coring; it is followed by additional required experimental procedures as outlined in Table 1.

This study introduces a novel approach that offers an alternative option to bypass the time-consuming experimental setup and associated delays. In order to illustrate the procedural steps of this study, a diagram depicting the process is created and presented in Fig. 1 . By utilizing base parameters to determine the quality of coal samples, independent of gas content operations, this approach widens the prediction perspective and overcomes the challenges posed by discontinuities and other factors affecting the non-homogeneity of coal gas distribution in underground systems. Additionally, the similarities between coal and shale reserves allow for the application of the same methodology in source rock exploitation for unconventional oil and gas systems, such as shale oil and shale gas.

2. Geological system and assumptions

This section provides a description of the geological system in Kınık coalfield. The coalfield is located on the border between Izmir and Manisa cities in Western Anatolia. Kınık coalfield belongs to the Soma

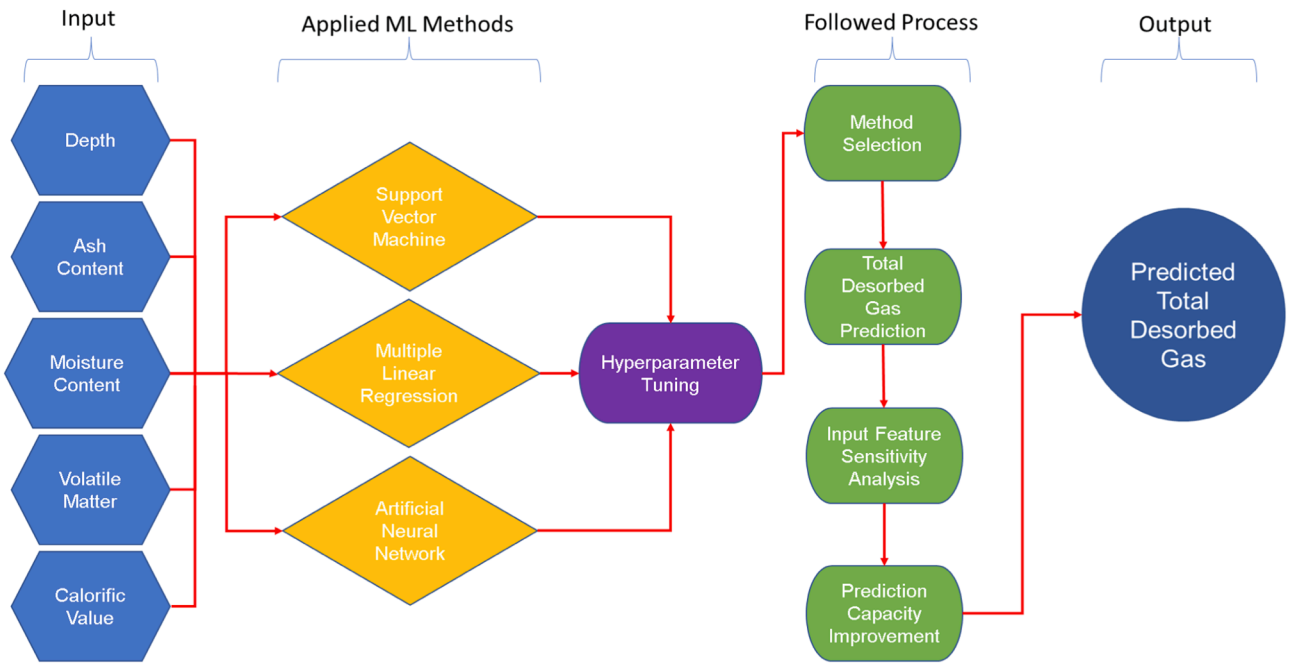


Fig. 1. Investigated Method diagram for prediction of coal gas content.

Coal basin, which has been exploited with open pit coal mines since 1950. The mining activities began with open pit mines and expanded through underground mining operations to extract additional proven reserves lying beneath the Soma region. According to estimates, Kınık coalfield has Turkey's deepest underground coal mine and is estimated to have 250 million tonnes of coal reserves [16]. The investigated area is well described by Esen et al. (2020) and is shown in Fig. 2.

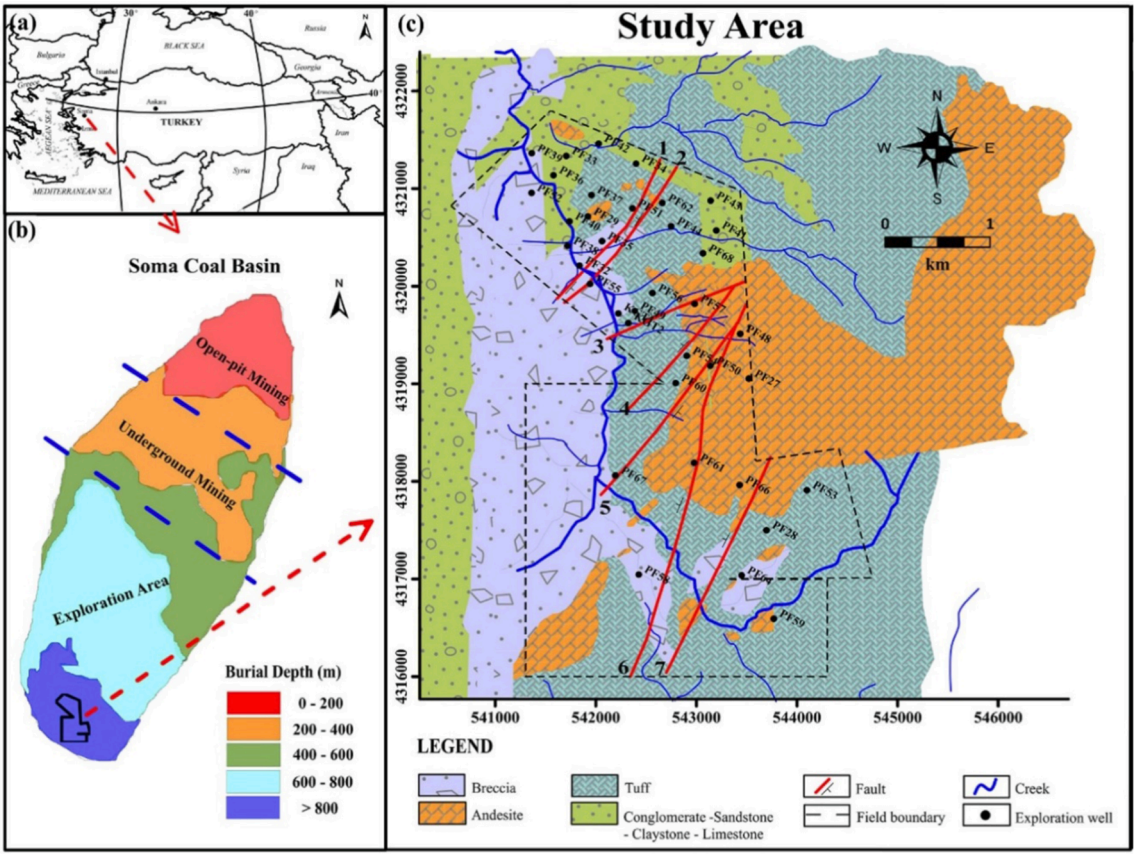


Fig. 2. (a) Soma Coalfield location in Turkey. (b) Location of Exploitation area in Soma Coalfield Basin. (c) Geological structures in studied area retrieved from Esen [16].

2.1. Geological system

The Soma coal basin stretches in a North-East to South-West direction and the open pit mining activities began at the North section of the basin, starting from the surface outcrop of the coal seams. The mining method then shifted to underground mining, after additional coal reserves were discovered at various depths [17]. The Kınık coalfield is mainly composed of Mesozoic carbonates and Miocene volcanic rocks, with two significant formations for coal production: the Soma and Deniş formations. The Deniş formation is relatively shallow and hosts the kP1 coal seam [18], while the Soma formation is deeper and contains two coal seams, kM2 and kM3. The stratigraphy of the Kınık coalfield is presented in Fig. 3. The Soma formation is a geological formation that originated in the lower-middle Miocene period. It has a thickness that ranges between 50 and 100 m and extends over a width of 4 km and a length of 25 km.

For this study, the data from two coal seams, kP1 and kM2, were used. Unfortunately, there was a lack of reliable core sample and gas content data available for the kM3 coal seam, which has a relatively small thickness compared to kP1 and kM2. Therefore, only 24 data points from kP1 and 60 data points from kM2 were analyzed in this study.

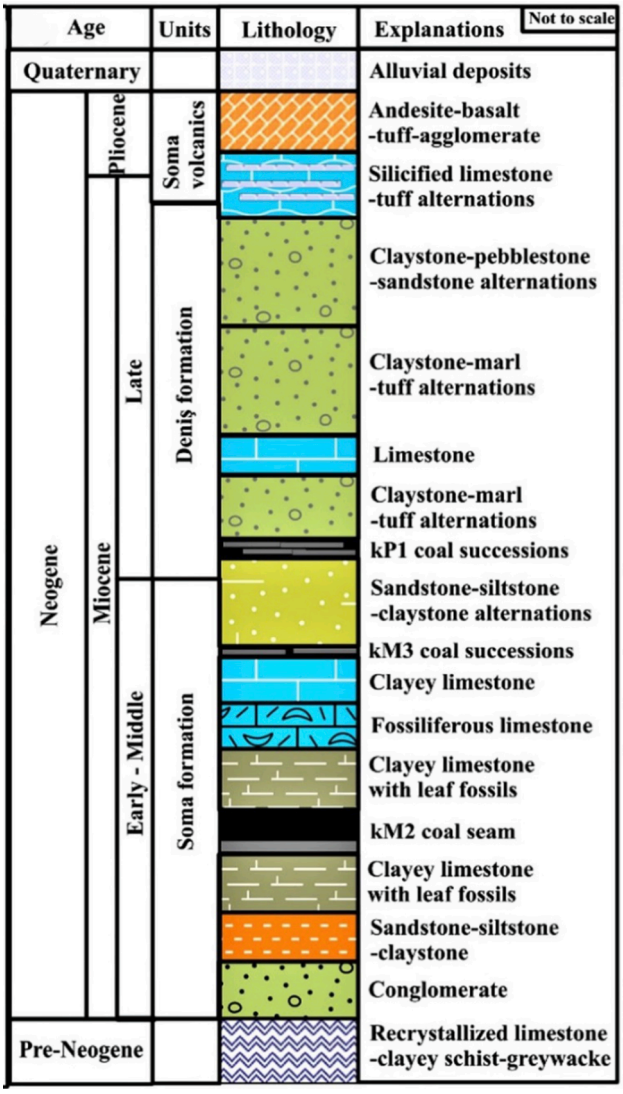


Fig. 3. Stratigraphic column of the Soma Basin [18].

2.2. Model assumptions

This study aims to utilize Machine Learning (ML) techniques to predict the gas content of coal samples in the Kınık coalfield. To achieve this, certain assumptions were made while processing the data points collected from kP1 and kM2 coal seams. It is assumed that the underground rock and coal seams are homogenous in lateral direction, and the fault system is underestimated due to the complexity of capturing and storage capacity of coal. Even if the fault and crack system allow for gas diffusion from the coal seam to upper cap rock via faults, the system is considered homogenous. It is also assumed that the solid rock and coal seams are always in local thermal equilibrium and the system is in equilibrium in constant pressure. Furthermore, it is assumed that coal seam maturation is complete, and there will be no further alteration. The datasets used in this study are based on five features, namely moisture, ash, volatile matter contents, depth, and calorific value of the given data points, with the aim of creating a general prediction algorithm for the Kınık coalfield.

3. Machine learning methods and results

In this section, the ML methods employed for predicting the total desorbed gas content in Kınık coalfield are presented. Due to the complex nature of the problem, ML offers a wide range of solutions where analytical methods fall short [19]. However, there is no established framework for ML-based prediction of coal gas content. Collecting data from underground systems is a costly and time-consuming process, and the variety of data collection techniques leads to additional expenses [20]. As a result, the number of drilling expeditions is limited, which, in turn, restricts core or data sampling. Given the scarcity of data and the high prediction capacity required for coal gas content, ML methods were considered. The selection criterion for the ML methods was the ability to handle relatively small datasets for evaluating complex underground systems. The Kınık coalfield dataset comprises 84 core sample analyses collected from kP1 and kM2 coal seams. Since there are no established guidelines for preprocessing the dataset into training, validation, and testing sets, the boundary is determined by the user to maximize the use of data. Typically, ML methods prefer an 80% training, 20% validation, and testing split if the dataset is sufficiently large [21]. The dataset consists of 84 lines of data with five different features, resulting in a total of 420 data points used to represent the Kınık coalfield. The dataset was split into 90% training and 10% testing sets. This study assumes that there is gas in each pore encountered within coal samples. Thus, the prediction algorithm is connected to the amount of gas rather than its productibility or classification as present or absent. The classification problem of gas content existence in coal samples was not considered as coal has primary porosity that does not diffuse bare gas in one-atmosphere laboratory pressure and secondary porosities filled with trapped gas content [12]. As a result, classification algorithms were not used in this study. Instead, three supervised learning-based regression methods were chosen to predict the total desorbed gas content in the Kınık coalfield.

3.1. Support vector machine

SVM is a powerful tool for the use of both classification and regression problems. In the case of the investigated dataset, a separating hyperplane is used in SVM classification to divide the data into two sections. The objective of the separation plane is to create two class systems from multi-dimensional data [22], which can be used to construct a model for prediction purposes.

Fig. 4 displays the two classes of SVM that are separated by a hyperplane, with support vectors located closest to the hyperplane. The margin between the hyperplane and the support vectors is computed using the Euclidean distance. The hyperplane is located in such a way as to maximize the classification accuracy, and its location is optimized

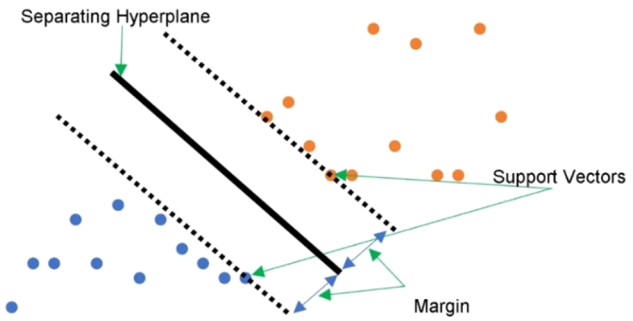


Fig. 4. Linear Classification.

accordingly [23].

The linear classifier can be written as described by Ben-Hur and Weston [24],

$$f(x) = w^{T}x + b. \tag{1}$$

In the linear classifier equations (Eq. (1), the inputs "x" and input weights "w^{T}" are combined with the bias term "b". However, in the nonlinear system, the equations are modified by introducing nonlinear parameters represented by the symbol "ϕ". Therefore, the modified equations are given by:

$$f(x) = w^{T}\phi(x) + b. \tag{2}$$

Simplification of nonlinear system shown in Eq. 2 via kernel function term were explained as,

$$\phi = \frac{1}{2}\|w^{T}w\| + C\sum_{i}\zeta_{i}. \tag{3}$$

The nonlinear parameters in Eq. (3) were expanded with the penalty parameter "C" and error "ζ". To simplify the nonlinear system, kernel functions were used in SVM classification to reduce the dimensionality from a nonlinear to linear system. In this study, the radial basis kernel function (RBF) was employed, as shown below:

$$K(x, x') = \exp\left(-r\|x - x'\|^{2}\right). \tag{4}$$

Eq. (4) contains the term "r," which is a radius that controls the kernel function. However, since this study focuses on predicting the total desorbed gas content, using support vector classification on the large number of unique data sets would result in an unreasonably large number of support vectors. Therefore, using a classification approach with floating-point numbers would be insufficient for predicting gas content. As a result, SVM regression was used, with a loss function built upon the SVM classification approach [25].

In this study, the dependent variable "y" represents the total desorbed gas content. The epsilon "ε" is a type of loss function that acts as a bridge between the transition from classification to regression, as shown in Eqs. (5).

$$LossFunction_{\varepsilon}(y) = \begin{cases} 0, |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, otherwise \end{cases} \tag{5}$$

The error and penalty parameters are working outside of the epsilon boundaries in Fig. 5 in respect to provide smooth transition between classification and regression. Overall, Support Vector Regression is using the Eq. 6 for the prediction performance [27].

$$f(x) = \left(w^{T}\phi(x) + b\right)K(x, x') \tag{6}$$

3.2. Multiple linear regression

Multiple linear regression (MLR) is a modeling technique that involves minimizing the sum of squared differences between the predicted (dependent) variable and multiple input (independent) variables using linear approximation [28]. The calculation for MLR involves the input variables denoted as Input "x_n", their corresponding weights as "w_n", the error term as "ζ", and the output relationship as shown in Eq. (7),

$$y = w_0 + w_1 \times x_1 + w_2 \times x_2 + \cdots + w_n \times x_n + \zeta. \tag{7}$$

In the MLR calculation the least square principle is using in order to minimize the absolute error between predicted and measured dependent variables as shown in Eq. 8,

$$LeastSquareError, \zeta = \sum_{i=1}^{n} (y - w_i \times x_i)' \times (y - w_i \times x_i) \tag{8}$$

3.3. Artificial neural network

Neuron cells in human body has inspired the ML method for the building of the Artificial Neural Network (ANN) method [29]. The basic concept of ANN involves the input and output relationship between two neurons through nodes and their corresponding weights that contribute to the final output. The ANN method that used in this study visualized in Figs. 6 and 7.

The ANN environment depicted in Fig. 6 includes variables such as input "x", bias "b", weight "w", and output "y". However, this environment can be extended by introducing the hidden layer logic. Unlike multiple linear regression, the ANN simultaneously uses inputs and their weights to take advantage of the different features' effect on each weight. In order to achieve this, the population of the node is represented in the hidden layer shown in Fig. 7.

The hidden layer in an ANN serves the purpose of creating connec-

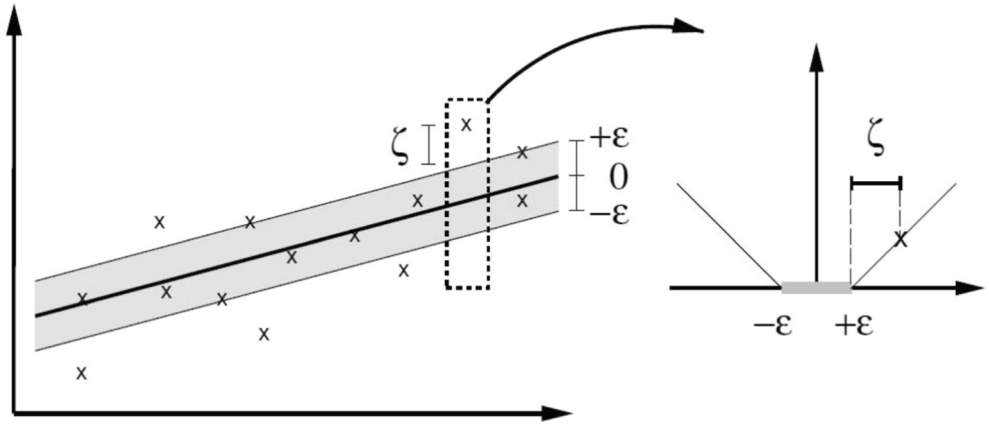


Fig. 5. Epsilon Insensitive Loss Function and Slack Variables retrieved from Smola and Schölkopf [26].

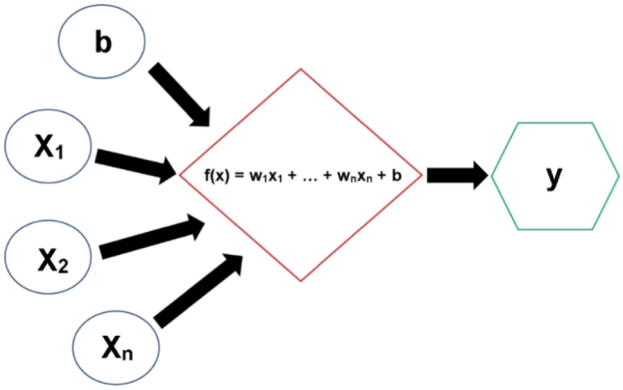


Fig. 6. Artificial Neural Network structure: The perceptron.

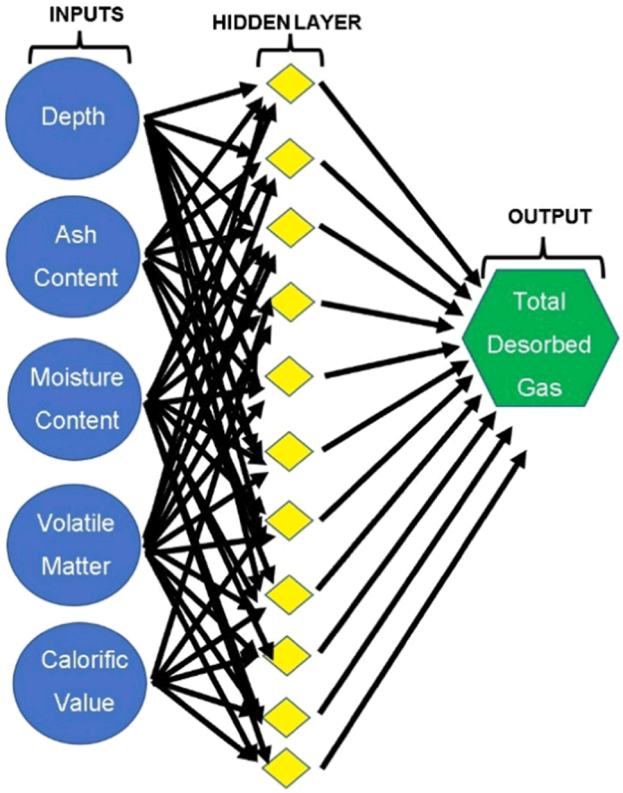


Fig. 7. Hidden layer demonstration of Artificial Neural Network *Results of the solutions.*

tions between inputs and outputs by introducing a population of nodes. This allows for more complex relationships between features to be captured and provides the flexibility to add multiple hidden layers. The number of nodes in a hidden layer determines the complexity of the model, and the weights between the nodes and inputs are learned during training. Each node in the hidden layer receives a signal from the inputs, which is modulated by a bias term and an activation function. The activation function determines whether the node should be activated or not based on the input signal and the boundaries specified by the function.

$$h_j = \sum_{i=1}^{n} w_{ij} x_i + b_j. \tag{9}$$

Eq. 9 shows the relationship between the hidden layer "h_j", input "x", bias "b", number of inputs "i", and number of hidden layers "j". In order to accumulate the node contributions, "Tanh" activation function [29] was used in this study, which is expressed in Eq. 10 as follows:

$$F_{tanh}(h_j) = \frac{1 - \exp(-2h_j)}{1 + \exp(-2h_j)}. \tag{10}$$

After the output calculations, ANN uses the least square (LS) method to calculate the error between the predicted and measured data output, as shown in Eq. 8. Then, it backpropagates to adjust the determined weight by nodes and hidden layer to minimize the error. In order to control the backpropagation, a learning rate is implemented in the backpropagation [30] sequence, as shown in Eq. 11.

$$w \leftarrow w + a(y - \hat{y})x. \tag{11}$$

Predicted "y ^" and measured "y" output variable difference in LS error was combined with input "x" and learning rate "a" in order to correct the weight "w" via backpropagation approach in Eq. 11.

The aim of this study was to improve the accuracy and efficiency of predicting the total desorbed gas content of coal, using machine learning (ML) methods. The data used in the study were collected from various locations and depths in the Kınık coalfield, and two types of visualizations were produced to evaluate the performance of the ML methods. The first visualization was the regression score "$r^{2\prime\prime}$, which indicates the model's success in predicting the total desorbed gas content of the entire coalfield using the input data. The second visualization was the error calculation for each data point. To build models for the Kınık coalfield, two ML methods were used: support vector machine (SVM) and artificial neural network (ANN). The hyper parameters for each method were optimized using the GridSearchCV algorithm, as shown in Tables 2 and 3.

There are two methods used for the parameter optimization which were SVM and ANN. Both methods use the least square error for weight optimization, and the errors are limited by the "Tolerance" constant in SVM, while ANN uses "lbfgs" to solve weight correction with a maximum iteration constraint and "alpha" for normalization of different types of features in the given dataset. Among the four activation functions available in the "Sklearn Library" for Artificial Neural Network (ANN) methods, the "Tanh" function was specifically chosen for this study using the GridSearchCV optimization algorithm. The selection was made after careful investigation of various activation functions, as it was observed that linear activation functions are not suitable for addressing the non-linear problem of coal gas content predictions. The "Tanh" activation function was preferred because it exhibits continuity and differentiability within the range of −1 and 1, with a centered value of 0. In comparison to other activation functions, "Tanh" has a steeper and less restrictive behavior when transitioning between activated (1) and deactivated (−1) gradients [31]. Furthermore, the success of the "Tanh" function can be attributed to the interdependence and bilateral effects of the features within the neural network. These features are often challenging to differentiate from one another while contributing to the connections between neurons. The different gradient characteristics of the "Tanh" function enable the analysis of all neuron feedings in the nodes, resulting in more efficient error handling and improved output accuracy.

In all predictions with ML methods, the random state function was used to fix the training and test datasets. ANN was controlled with only one hidden layer, shown in Fig. 7, to ensure acceptable computational cost. This study considered three types of machine learning (ML) methods, two of which required hyperparameter optimization. The optimization process involved setting various parameters such as data point evaluations, neighborhood functions, distance metrics between

Table 2
SVM tunned hyperparameters.

Kernel Function	Tolerance	Epsilon	C
RBF	0.1	0.1	325

Table 3
ANN tunned hyperparameters.

Hidden Layer Size	Maximum Iteration	Activation Function	Tolerance	Random State	Solver	Alpha	Learning Rate
375	10,000	Tanh	0.001	7	lbfgs	1.00E-04	0.001

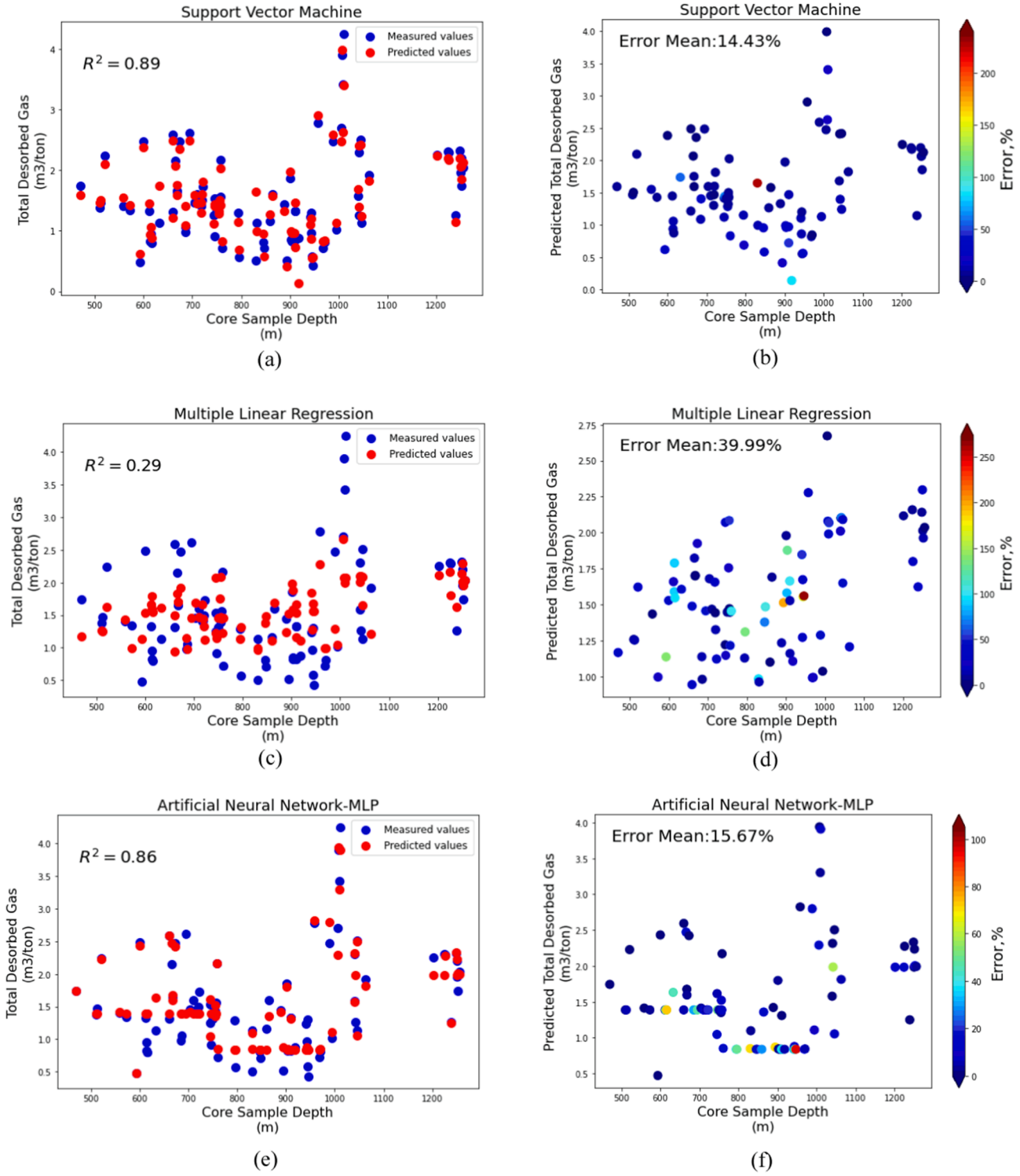


Fig. 8. (a) Regression score of Kınık Coalfield via SVM, (b) Error map of predicted gas content of individual data points in Kınık Coalfield via SVM, (c) Regression score of Kınık Coalfield via MLR, (d) Error map of predicted gas content of individual data points in Kınık Coalfield via MLR, (e) Regression score of Kınık Coalfield via ANN, (f) Error map of predicted gas content of individual data points in Kınık Coalfield via ANN.

data points, tolerance values for weight optimization, activation functions for modeling relationships, and the number of iterations for achieving the final model configuration. To perform this hyperparameter optimization, the GridSearchCV algorithm was employed. This algorithm systematically evaluates the model's performance using different combinations of hyperparameters. It measures the model score during the training phase and identifies the best-fit combination of hyperparameters for the selected ML methods. The algorithm searches through all the given hyperparameter combinations and selects the one that yields the highest performance. This best-fit combination is then utilized to construct the actual model and make predictions on the target variables. The results of the predictions using ANN ML methods are shown in Fig. 8.

Fig. 8 presents the analysis of the field dataset, consisting of 84 core samples from the Kınık coalfield, to evaluate the performance of the machine learning methods. Measured and predicted data are depicted in Figs. 8a, 8c, and 8d. In this analysis, 10% of the data was excluded and the remaining 90% was used to train the system. Subsequently, the entire dataset was re-predicted to assess the effectiveness of the selected method. Additionally, Figs. 8b, 8d, and 8f illustrate the error map of the predicted field data. The color bar on the right side of the figure represents the absolute error for each individual data point in the field data prediction. The color gradient ranges from blue to red, with red indicating the worst prediction characterized by the highest absolute error. It is important to note that Fig. 8 serves the purpose of validation. While it is generally recommended in the literature to avoid using the same ML learning data for both training and predicting the entire dataset of the system, this study aims to pioneer new approaches in order to establish a foundation for coal gas content prediction. Thus, unconventionally, the given system employs the same dataset to define the measured gas and predict the gas content, with only 10% of the data excluded for testing purposes. Careful consideration was given to selecting the test data points in a manner that would not disrupt the overall reflection of the system. These test data points were chosen based on their proximity to neighboring points and their positioning within the boundaries of the dataset. It was desired to have 9 target variables from the core samples in the system, ensuring tolerability during system definition and maximizing the presence of neighboring data points for improved prediction performance.

4. Sensitivity analysis of the input features

In this section, a sensitivity study was carried out to examine the impact of the input features on the prediction performance of the gas content. The aim was to investigate how the features of the coal samples affect the gas content prediction through training, regression, and error scores of ML methods. Thus, the preliminary analysis performed in Chapter 3 was evaluated to determine which ML algorithm is suitable for feature performance analysis.

Chapter 4 is dedicated to the comprehensive explanation of the measured and target variable system using ML methods, sensitivity analysis of input features, and the computational capabilities of the employed ML methods. Fig. 9 consists of three columns illustrating the sequence of the used machine learning methods in terms of the measured and simulated target variables. The model score, which assesses the performance of the measured target variables, is employed to evaluate the success of the ML methods on the training data. The succession of the model score can be measured, with a perfect score of 1 indicating optimal performance. Fig. 9 displays the training performance, regression score, and mean squared error score of each ML method for the Kınık coalfield dataset. Based on the training capacity and the slightly better performance in the regression and error scores, it can be concluded that SVM is the most suitable method for predicting the total desorbed gas content in the Kınık coalfield. Additionally, the computational cost is significantly lower with the simpler SVM design compared to the ANN method, as illustrated in Fig. 10. Therefore, the sensitivity analysis was further conducted using the SVM method.

During the parametric study, five features were considered, namely:

- Depth,m
- Moisture content,%
- Ash content,%
- Volatile Matter content,%
- Calorific Value,MJ/kg

Fig. 11 compares the features of the Kınık coalfield dataset and shows a clear relationship between ash content and calorific value with respect to the total desorbed gas content. To verify the pair plot analysis of the dataset and determine the total desorbed gas prediction capacities of each feature, SVM regression method was used to analyze each feature independently in Fig. 12.

Although the model training and regression scores for "Calorific Value" were low in Fig. 12, it showed the most comprehensive scanning of the Kınık coalfield desorbed gas distribution along with "Ash Content". Moreover, the "Depth" of the used samples contributed to the analysis of the total desorbed gas content distribution. On the other hand, "Moisture Content" and "Volmat, volatile matter content" were found to be the least effective features for gas prediction capacity.

In Fig. 13, the mean squared errors were analyzed to further support the regression scores in Fig. 12b in terms of feature prediction performance. It was observed that each feature contributed differently to the error calculations. Hence, it is not feasible to exclude any particular feature as they all play a role in accurately characterizing the investigated coalfield. For instance, although the "Moisture" content had the highest error in prediction, it is a crucial part of the environment that creates the coal system along with other given parameters. Therefore, it cannot be disregarded. The purpose of ML approaches is to develop an interconnected comprehension of the unique features. One of the findings obtained from Fig. 13 is that it can be used for future analysis in outlier detection in sufficiently large datasets to obtain a better understanding of individual data points.

5. Prediction capacity of optimized ML for test data points

The main objective of this research is to develop a reliable ML-based method for estimating the in-situ gas content. To achieve this goal, a

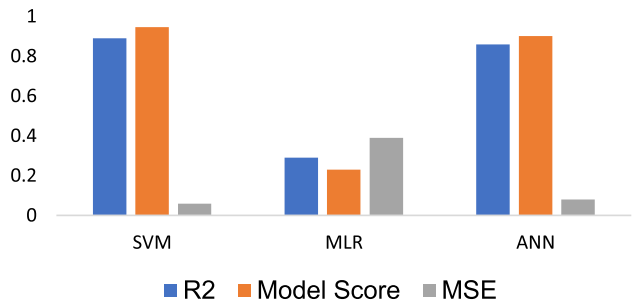


Fig. 9. ML Methods Scores for Kınık coalfield analysis.

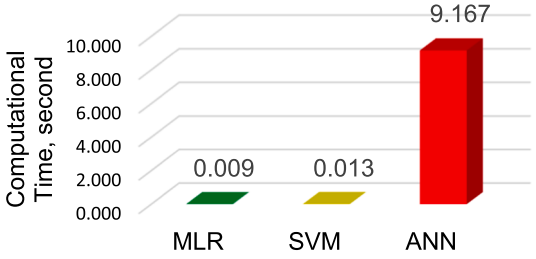


Fig. 10. ML method computational cost evaluation for Kınık coalfield dataset.

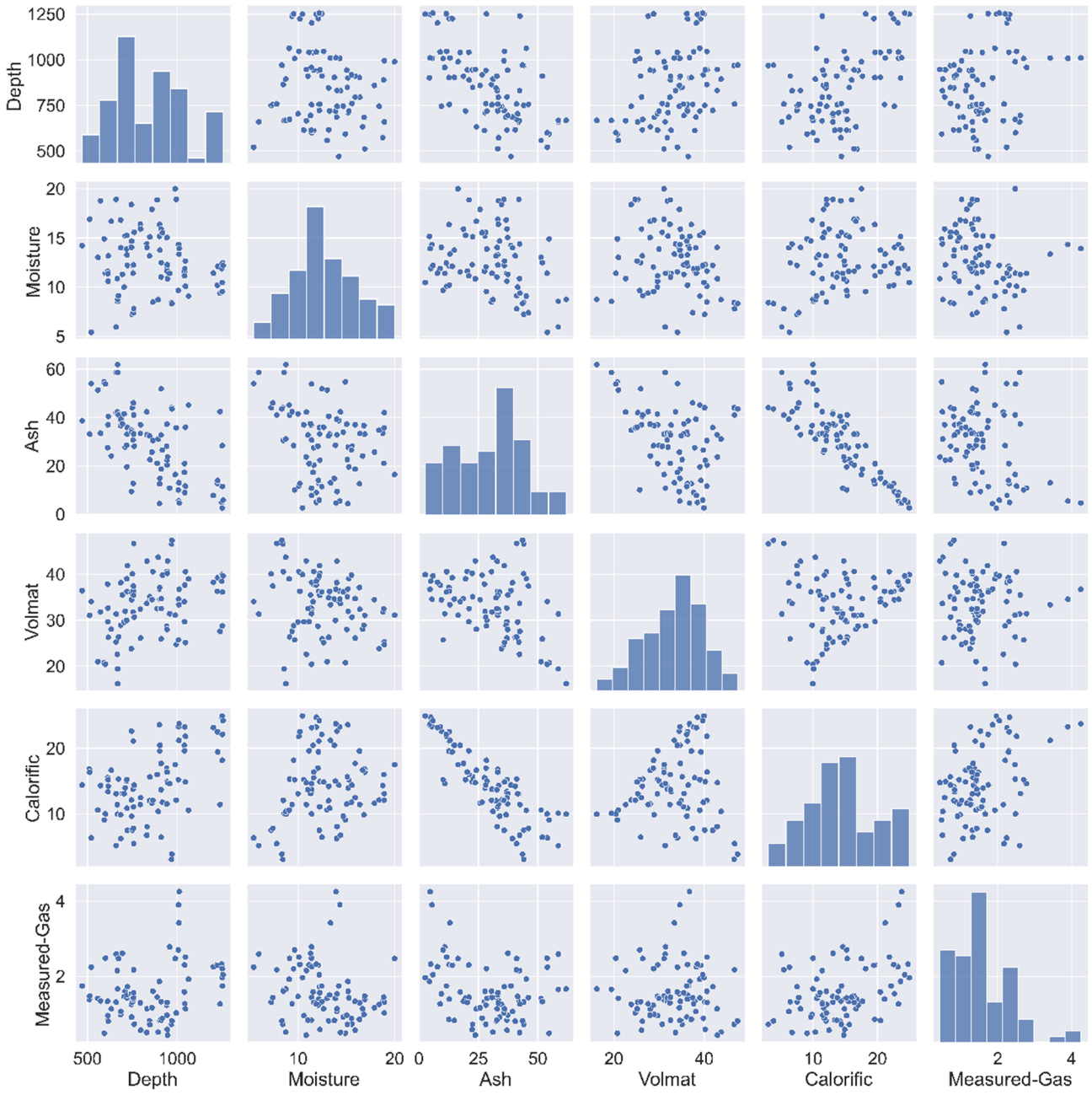


Fig. 11. Kınık Coalfield features pair plot.

support vector machine regression model was constructed due to its superior prediction performance. The dataset was divided into two sets, with 10% of the data points reserved for testing and the remaining 90% used for training the model. Accordingly, 9 data points out of the total 84 data points were set aside for testing purposes during the model building phase.

The test data points were selected based on their proximity to other data points in the Kınık coalfield dataset, in order to obtain a better understanding of the desorbed gas distribution network. Specifically, data points with the most neighbors were chosen as test points. This selection criterion was chosen because test points surrounded by dense training data points tend to yield better prediction results. Additionally, the selected test points were chosen based on their relevance to the field definition for desorbed gas distribution. Overall, this approach was designed to maximize the accuracy of the model and ensure that it would perform well on new, unseen data points.

The initial model construction and analysis showed insufficient results, as seen in Fig. 15. The regression score indicated a similarity of 0.31 between the measured and predicted test data. To further investigate the data and sort it by depth, a test data analysis was performed using SVM regression, and the results are presented in Table 4.

The SVM model constructed in Chapter 3 was used to analyze the test dataset presented in Table 4 individually. The aim of the individual analysis is to expand the training dataset along with the rest of test data points while one at a time continue to being tested. New testing sequence with one data out of 84 point was resulted as,

By analyzing each designated test point individually with the same SVM model used in Chapter 3, an improved regression score was achieved, as depicted in Fig. 16. This approach of analyzing each data point individually resulted in a significant improvement in the regression results from 0.31 to 0.67 compared to the analysis of the 9-point test dataset. Furthermore, the mean absolute error for the tested data points

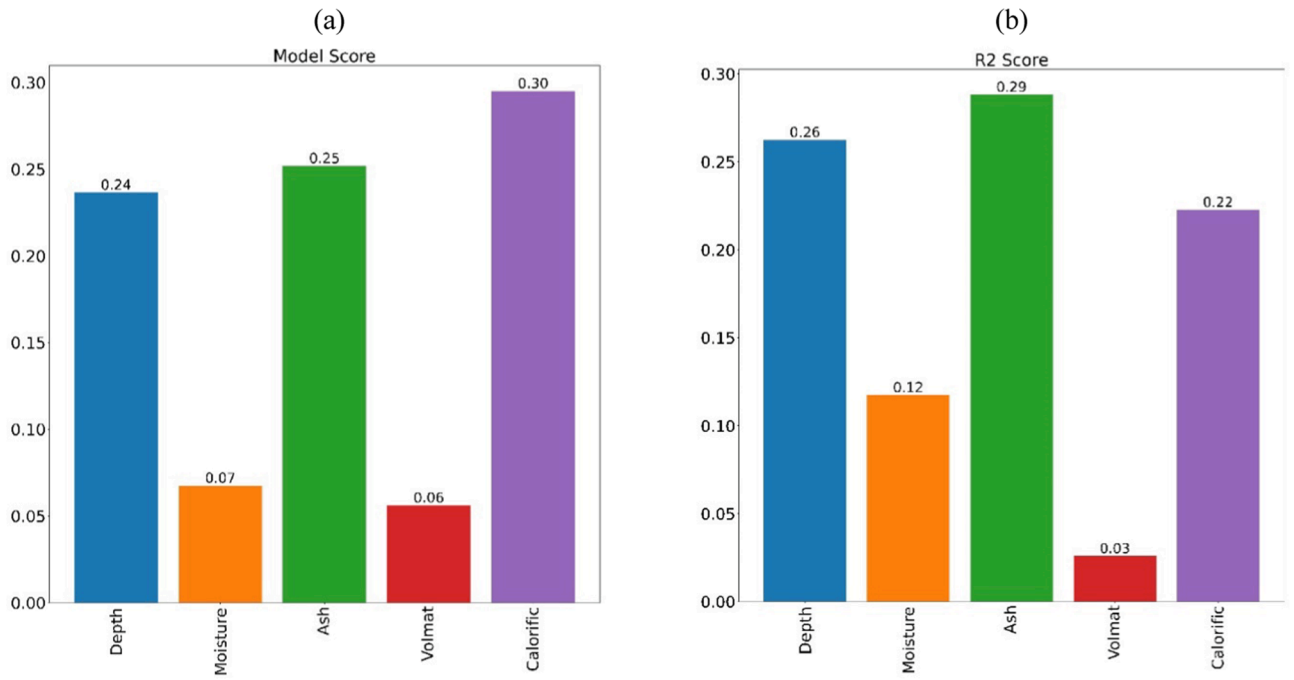


Fig. 12. (a) Model training score of Kınık Coalfield features via SVM, (b) Regression score of the Kınık Coalfield via SVM.

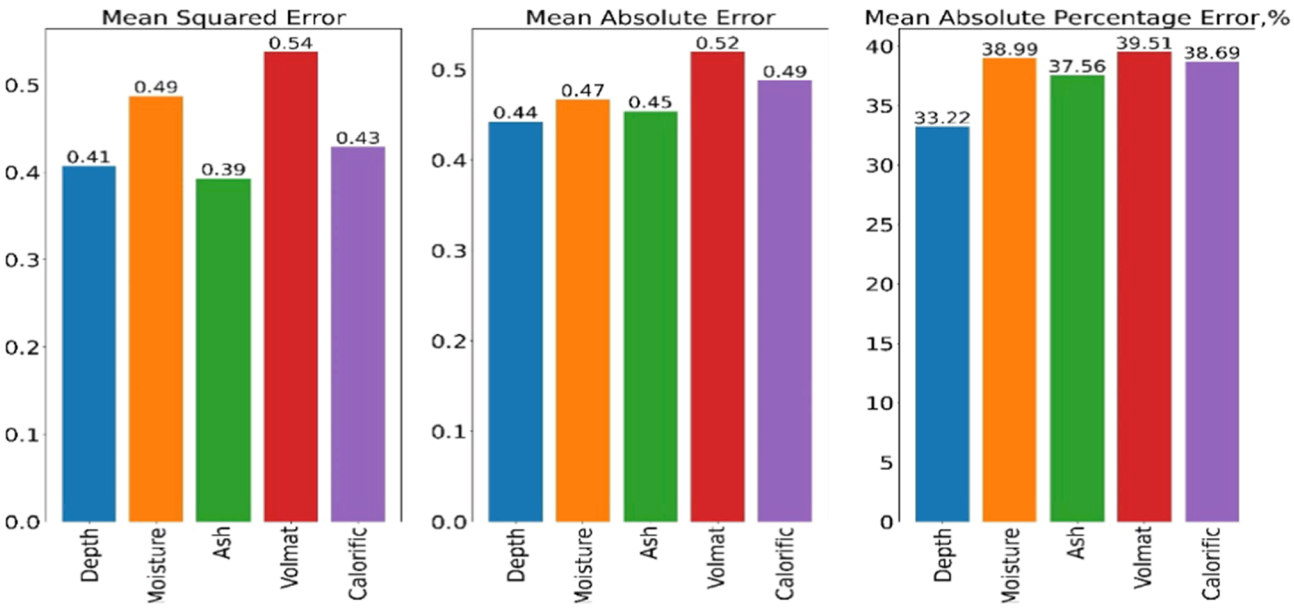


Fig. 13. Error scores of the Kınık Coalfield features respect to SVM regression analysis.

Table 4
Test data prediction results with SVM model.

Index	Depth, m	Measured Total Desorbed Gas, m^3/t	Predicted Total Desorbed Gas, m^3/t	Absolute Error, %
12	632.6	1.13	1.74	53.77
20	685.65	1.06	1.4	32.34
30	746.03	0.91	1.44	58.13
32	753.65	1.43	1.28	10.18
40	830.1	0.5	1.65	229.55
52	910.2	1.33	0.72	45.73
54	917.75	0.88	0.14	84.33
62	970.3	0.81	0.84	4.32
65	1006.1	2.7	2.47	8.38

was reduced from 58.53% to 32.47%.

This study was conducted using 84 core samples from the Kınık coalfield. To further investigate the generalization of the constructed ML method using SVM regression, additional core samples were collected from different coalfields with various types of coal samples. However, these new core samples did not have calorific values, so the analysis was limited to features such as depth, moisture content, ash content, and volatile matter content. This was done in order to make use of the additional field data and improve the generalization of the constructed ML method for various types of coal samples.

The inclusion of new data entries has significantly improved the regression score for individual data points compared to the previous test dataset from the Kınık coalfield. The prediction performance for a single data point out of the 368 data points has a regression score of 0.83,

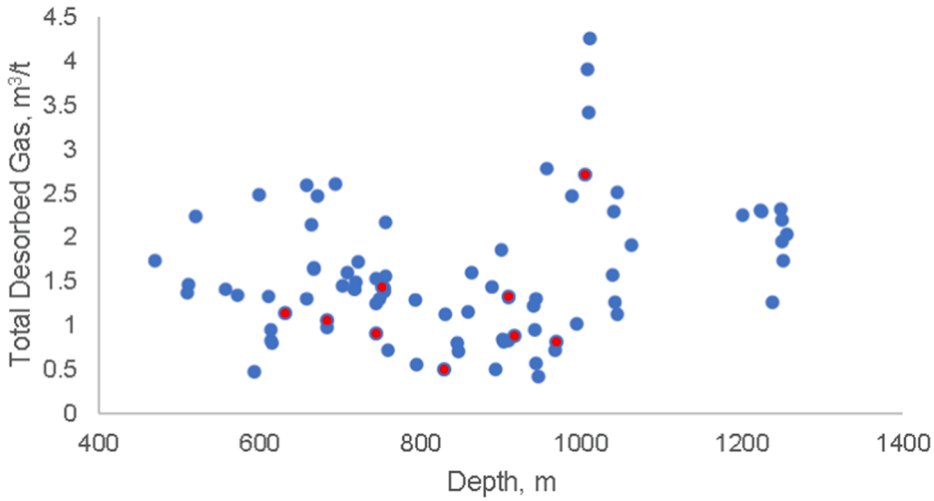


Fig. 14. Selected train (Blue dots) and test (Red dots) data from Kınık coalfield dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which is higher than the regression score of 0.67 obtained for the 84 data point analysis. This demonstrates that a larger dataset results in better prediction performance. However, the absence of the "calorific value" feature, which is known to contribute significantly to the prediction performance, has weakened the ability of error calculation.

In Chapter 3, ML methods were evaluated for their success in predicting gas content in the field, using a dataset where 10% of the data points were excluded. Based on the comparison of regression scores, model scores, and mean absolute error scores, SVM was selected for further analysis of coal gas content. In Chapter 5, the analysis was conducted using 84 core samples in three stages. Initially, the dataset was divided into 10% test data (9 target data points) and 90% training data (75 data points). The results for the test data in the first configuration were presented in Fig. 15, but the desired regression score was not achieved. To enhance the training mechanism, the second configuration involved testing one out of 84 samples (as the target) while using the remaining 83 samples as training data. The analysis was conducted individually for each target data point, resulting in an improved regression score of 0.67 (compared to 0.32 in the first configuration), as shown in Fig. 16. In the final configuration, additional data from different coal reserves were included to further improve the target prediction capacity as shown in Fig. 17. However, this new dataset did not include the calorific value feature, and only four features (depth, moisture content, ash content, and volatile matter content) were utilized. Although the number of data points increased significantly from 84 to 368, the absence of a crucial feature led to an increase in mean absolute error from 32.47% to 60.58% (compared to 58.53% in the first configuration). It is important to note that both the variety and size of the input dataset have a significant impact on the prediction capacity (see Figs. 16 and 17).

6. Discussion and future study

The conventional methods for determining coal gas content, as presented in Table 1, involve physical labor and add to the workload of regular exploration tasks. Moreover, these methods require complex experimental procedures and technical expertise. In contrast, this study focuses on fundamental coal sample features that are essential for every coal reserve, eliminating the need for labor-intensive and specialized techniques.

Furthermore, gas content determination can vary depending on the operator and the time intervals of measurements in both direct and indirect methods. This variability often leads to overestimation or underestimation of gas content results. This study offers a solution by providing a time and operator-independent approach for consistent gas content prediction. The results show promising agreements when compared in terms of regression score and absolute error, enhancing the reliability of the predictions.

This research explores the application of machine learning (ML) techniques, which enables the examination of gas content at each

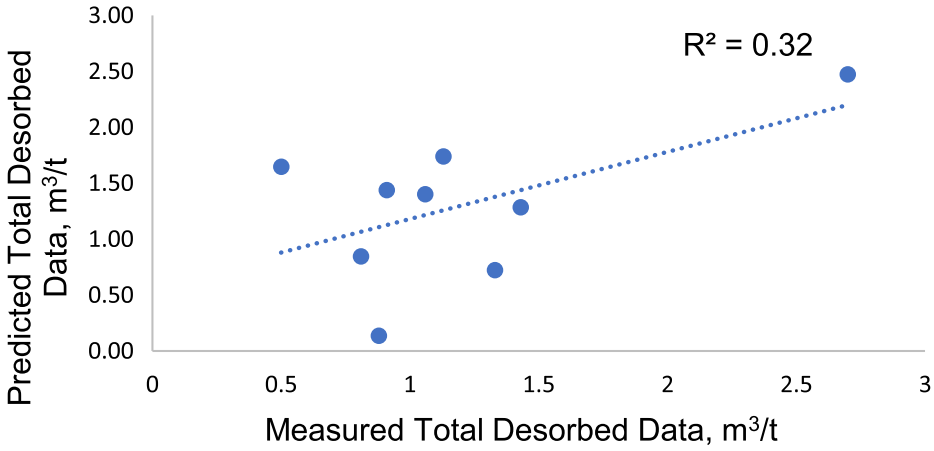


Fig. 15. SVM regression analysis for the test data prediction capacity.

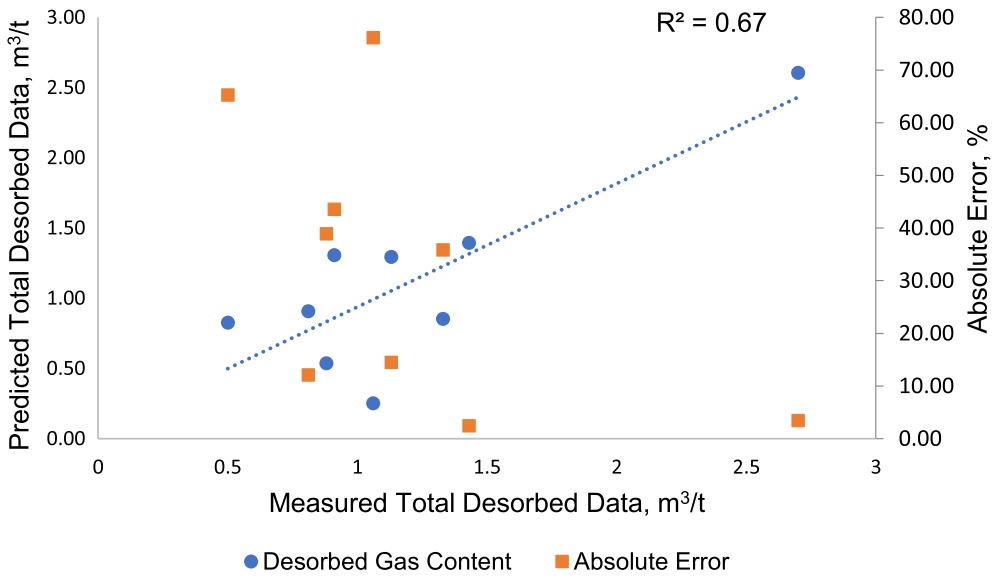


Fig. 16. SVM regression analysis for individual tested points prediction capacity.

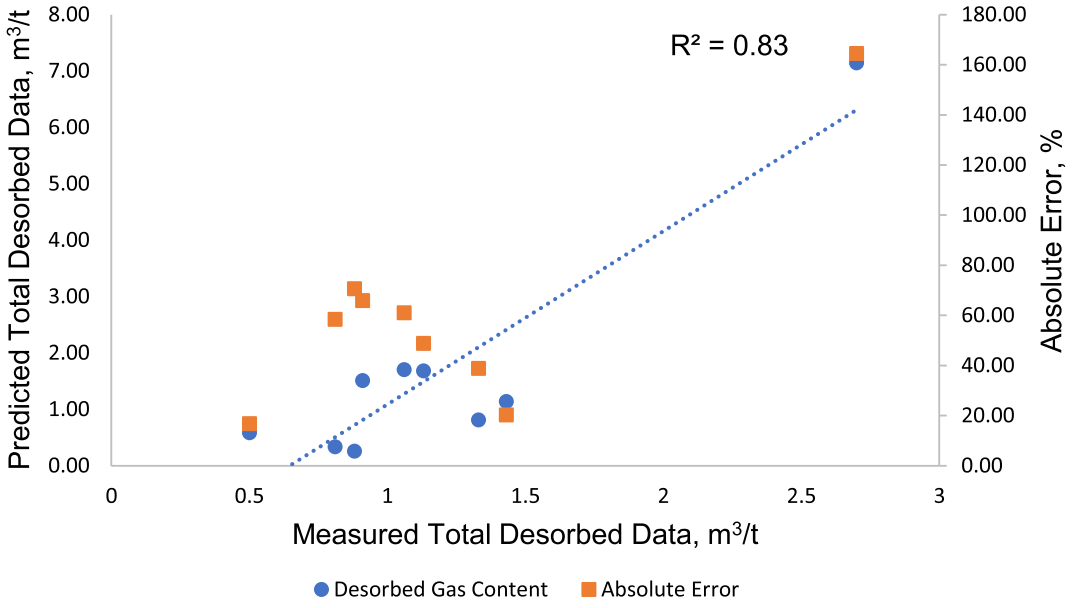


Fig. 17. SVM regression analysis for individual tested points prediction capacity with New Data Entries.

individual depth. Unlike traditional approaches that rely on maximum or weighted gas content measurements for reserve evaluations, ML methods provide a wider scope for evaluating underground systems. Among the ML methods tested, Support Vector Machine (SVM) proves to be more suitable compared to Artificial Neural Network (ANN) and Multiple Linear Regression (MLR). SVM utilizes planes to create prediction algorithms, while ANN involves complex relationships and MLR relies on linear regression with given features along straight lines. The design of the system is reciprocal, as coal systems are defined in planes for evaluation purposes. Consequently, SVM demonstrates better agreement in terms of regression performance and computational time, thus enhancing the overall prediction capacity.

The inclusion of coordinates and depths of coal samples allows for the creation of a 3D solution to assess gas content in a given area. This preliminary model resembles the gas distribution in geological systems, resembling layers in commercial integrated reservoir models. In this study, the design of the Artificial Neural Network (ANN) was kept simple for the purpose of comparison with SVM and MLR, considering the computational and optimization challenges associated with the hardware and interfaces used. However, there is significant potential for exploring more advanced Neural Network designs and alternative regression models to enhance prediction capacity, computational efficiency, and accuracy. Therefore, this innovative study utilizing ML techniques and feature combinations provides support for alternative methods in regression analysis within this field.

7. Summary and conclusions

Determining the gas content in coal is a difficult and challenging task due to its heterogeneous nature, and conventional methods are expensive and time-consuming. This creates a need for alternative methods that can predict gas content of coal accurately and efficiently. This study aimed to develop a data-driven machine learning model for predicting the total desorbed gas content of coal seams from various core samples using Support Vector Machine Regression. Compared to other methodologies, SVM was found to be more effective in predicting gas content of

coal. The model was constructed using a limited dataset from the Kınık coalfield and was found to be applicable to different types of coal samples and fields. The ultimate goal of this study was to predict the gas content of coal seams at various depths accurately, which is crucial for ventilation design in underground mining operations. The SVM model was sensitive to the input features, and the study determined the effect of each feature on gas content prediction. This study not only improves the accuracy of predicting gas content in coal but also paves the way for investigating underground energy sources with higher accuracy.

In summary, this study has yielded the following findings:

- Even with limited data points, ML models can be constructed to predict coal gas content, although their accuracy may be relatively weak.
- The unique structure and features of coal have a significant impact on the absorbed gas content.
- The number of training points and missing features greatly affect the reliability of predictions for specific points.
- Multiple featured datasets may have various scales that are difficult to deal with, requiring preprocessing such as standardization for SVM and normalization for ANN methods.
- The system designed in this study allows for better interpretation of gas content between boundaries, but new fields or outer boundary predictions will require alternative prediction models due to a lack of training data for those sections.
- Additional features, such as elemental, chemical, petrographic, and mineralogical compositions, can be incorporated to expand the number of features and build stronger prediction models.
- Data-driven models show good agreement with predictions from nonlinear and heterogeneous systems in low rank coal samples, suggesting that ML methods can replace conventional analytical approaches that yield insufficient models.
- Total desorbed gas content of coal has similar structures to shale oil and gas reservoirs, suggesting that ML methods can be applied to field investigations of coal and reproduced for shale reservoirs under certain circumstances and assumptions.

CRediT authorship contribution statement

Satuk Buğra Akdaş: Data curation, Writing- Original draft preparation, Visualization, Investigation, Software, Validation and Editing, Methodology, Conceptualization, Project Administration. **Abdullah Fisne:** Data curation, Project administration, Resources, Supervision.

Declaration of Competing Interest

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

Data availability

Dataset used to build model from Kınık region is shared in one of the references. At the end, model was expanded with confidential data. If it is required, new data can be shared without label.

Acknowledgments

This paper is a part of the first author's MSc. study at the Istanbul Technical University.

References

[1] Company BP. BP Statistical Review of World Energy. London: British Petroleum Co., London; 2021.
[2] Willing N. Capital. Capital Com Bel, 20 April 2022. [Online]. Available: https://capital.com/coal-price-forecast. [Accessed 20 April 2022].
[3] Guo Z, Zhao J, You Z, Li Y, Zhang S, Chen Y. Prediction of coalbed methane production based on deep learning. Energy 2021;230.
[4] Karacan CÖ, Ruiz F, Cotè M, Phipps S. Coal mine methane: A review of capture and utilization practices with benefits to mining safety and to greenhouse gas reduction. Int J Coal Geol 2011;86(2–3):121–56.
[5] Kovalchuk N, Hadjistassou C. Fathoming the mechanics of shale gas production at the microscale. J Nat Gas Sci Eng 2020;78.
[6] Fathi E, Tinni A, Akkutlu I. Correction to Klinkenberg slip theory for gas flow in nano-capillaries. Int J Coal Geol 2012;103:51–9.
[7] Gao L, Mastalerz M, Schimmelmann A. The origin of coalbed methane. In: Coal Bed Methane Theory and Applications, Elsevier; 2020. p. 1.
[8] EIA. Natural Gas. EIA, 21 April 2022. [Online]. Available: https://www.eia.gov/dnav/ng/hist/rngr52nus_1a.htm. [Accessed 21 April 2022].
[9] Zhou F, Xia T, Wang X, Zhang Y, Sun Y, Liu J. Recent developments in coal mine methane extraction and utilization in China: a review. J Nat Gas Sci Eng 2016;31:437–58.
[10] Tang Y, Gu F, Wu X, Ye H, Yu Y, Zhong M. Coalbed methane accumulation conditions and enrichment models of Walloon Coal measure in the Surat Basin, Australia. Nat Gas Ind B 2018;5(3):235–44.
[11] Saghafi A. Discussion on determination of gas content of coal and uncertainties of measurement. Int J Min Sci Technol 2017;27(5):741–8.
[12] Diamond W, Schatzel S. Measuring the gas content of coal: a review. Int J Coal Geol 1998;35(1–4):311–31.
[13] Xue S, Yuan L. The use of coal cuttings from underground boreholes to determine gas content of coal with direct desorption method. Int J Coal Geol 2017;174:1–7.
[14] I. 18871. Method of Determining Coalbed Methane Content. ISO (the International Organization for Standardization), Geneva, Switzerland; 2015.
[15] Hou X, Liu S, Zhu Y, Yang Y. Evaluation of gas contents for a multi-seam deep coalbed methane reservoir and their geological controls: In situ direct method versus indirect method. Fuel 2020;265.
[16] Esen O. Soma Kömür Havzası Kömür Damarlarının Gaz İçeriği,Gaz Depolama Kapasitesi ve Gaz Akış Özelliklerinin Araştırılması. Istanbul Technical University; 2021.
[17] Nebert K. Linyit içeren Soma Neojen bölgesi, Batı Anadolu. Maden Tetkik Arama Enst. Derg. 90(90) (1978) 20-70.
[18] Oskay R, Bechtel A, Karayiğit A. Mineralogy, petrography and organic geochemistry of Miocene coal seams in the Kınık coalfield (Soma Basin-Western Turkey): Insights into depositional environment and palaeovegetation. Int J Coal Geol 2019;210.
[19] Xu L, Wang B, Du X, Hong Y. Prediction method of mine gas emission based on complex neural work optimized by Wolf pack algorithm. Syst Sci Control Eng 2018;6(3):85–91.
[20] Wang L, Cheng L, Cheng Y, Liu S, Guo P, Jin K, et al. A new method for accurate and rapid measurement of underground coal seam gas content. J Nat Gas Sci Eng 2015;26:1388–98.
[21] Salazar J, Garland L, Ochoa J, Pyrcz MJ. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. J Pet Sci Eng 2022;209.
[22] Boswell D. Introduction to Support Vector Machines; 2022.
[23] Gunn S. Support vector machines for classification and regression. University of Southampton; 1998.
[24] Ben-Hur A, Weston J. A User's Guide to Support Vector Machines, Clifton, N.J.: Methods in molecular biology; 2010.
[25] Smola A. Regression estimation with support vector learning machines (Master's Thesis). Technische Universität München, Physik Department, München; 1996.
[26] Smola A. A tutorial on support vector regression (Technical report). Australia: RSISE; 2003.
[27] Awad M, Khanna R. Support vector regression. In: Efficient Learning Machines, Apress; 2015. p. 67–80.
[28] Montgomery D, Peck E, Vining G. Multiple Linear Regression. in Introduction to Linear Regression Analysis, Wiley; 2012. p. 67.
[29] Belyadi H, Haghighat A. Machine learning guide for oil and gas using python: a step-by-step breakdown with data, algorithms, codes, and applications. Elsevier; 2021.
[30] Noriega L. Multilayer Perceptron Tutorial. School of Computing: Staffordshire University; 2005.
[31] Sharma S, Sharma S, Athaiya A. Activation function in neural networks. Int J Eng Appl Sci Technol 2020;4(12):310–6.