IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 25, 2021, accepted December 6, 2021, date of publication December 8, 2021,
date of current version December 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3133889

Enhanced Optimization-Based Voting Classifier and Chained Multi-Objective Regressor for Effective Groundwater Resource Management

ATIF RIZWAN[1], ANAM NAWAZ KHAN[1], NAEEM IQBAL[1], (Member, IEEE), RASHID AHMAD[2,3], AND DO HYEUN KIM[4]

[1]Department of Computer Engineering, Jeju National University, Jeju-si, Jeju-do 63243, Republic of Korea
[2]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[3]Bigdata Research Center, Jeju National University, Jeju-si, Jeju-do 63243, Republic of Korea
[4]Department of Computer Engineering and Advanced Technology Research Institute, Jeju National University, Jeju-si, Jeju-do 63243, Republic of Korea

Corresponding author: Do Hyeun Kim (kimdh@jejunu.ac.kr)

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2021-0-00188, Open source development and standardization for AI enabled IoT platforms and interworking) and this research was supported by Energy Cloud R&D Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (2019M3F2A1073387), Any correspondence related to this paper should be addressed to DoHyeun Kim.

ABSTRACT Water is an essential source of life for every living thing, and drilling is the only source to gain water from underground. Different advanced technologies have been used to minimize the time factor and labor force. Along with technology to be used, some other factors are equally essential to be considered, like water level, the hardness level of the land, and the number of days spent on the whole process. The study proposed a weighted voting classifier based on Differential Evaluation (DE) to classify the regions with different soil colors and land layers. The weights are assigned to the candidate classifiers based on their performance for each class. For the assignment of the optimal weight, the DE optimization algorithm is used. Moreover, the study presents a chained multi-objective regression model to simultaneously predict the water level and total depth on different locations. The proposed work facilitates the drilling industry to increase the rate of penetration (ROP) by selecting the region with soft soil and land layer. The prediction of depth and water level allows the industry to estimate water levels in different areas at different depths. The dataset is provided by the research organization, which contains information of different drilling points. The results of the proposed weighted voting classifier are compared with the traditional machine learning models (kernel Naive Bayes, Gaussian SVM, Quadratic SVM, and Bilayered Neural network) and state of the art voting classifier in terms of precision, recall, and accuracy. Moreover, the proposed regression model is evaluated by well-known evaluation metrics, including Mean Absolute Error, Mean Square Error, and R2 score. Finally, the comparison verifies the effectiveness of the enhanced optimization-based classifier and multi-objective regressor.

INDEX TERMS Weighted voting classifier, multi objective regression, groundwater level, planning and risk assessment, water resource management.

I. INTRODUCTION

Efficient and Robust models are required in the field of city construction and drilling related problems. As we required labor force and time for the process of drilling. The process is also expensive in terms of cost and equipment to be used [1]. So, there should be a track or a way that should be followed when drilling process is started. Furthermore, the type of land in different areas and layers of land should be known to reduce the risks such as stuck pipe, formation fracturing, and lost circulation [2]. There are multiple soil colors on the different layers of the land. These soil colors play an important role in the process, when we know the pattern of soil colors that lead us to the water [3]. Moreover, the number of days and the maximum depth should also be known to plan the project and gain the water. By taking into account all these

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

problems in city construction, the proposed study predicts the soil color and layer on different areas and estimates the water level and depth.

Drilling is one of the important and only source to gain water from underground. The water has different underground levels, and the different layers of the land layers lead us to the water [4]. In this study, the soil color and land layer are predicted by using the proposed DE-based weighted voting classifier, and results are compared with some state-of-the-art ML algorithms and ensemble voting classifier [5]. In ensemble techniques, multiple classifiers are trained, and the knowledge of all classifiers is used to make the final decision [6]. As compared to the individual classifier, the combination of multiple classifiers could be more effective [7]. To get a good ensemble classifier, the candidate classifiers should be as strong as possible. The selection of the best member classifier is a concern of many researchers [8]. There are multiple ensemble techniques for classification problem including bagging, boosting and stacking [9]–[11]. For the multi-class problem, the combination of multiple classifiers can be more effective. In the dataset of drilling, we have ten different soil colors and eight rock layers of the land. The dataset is provided by a research organization, which contains each borehole's information on a certain location. Most of the features in the dataset are numeric continuous, and other features are encoded into the numeric value to make the calculations easy.

The dataset contains different soil colors and land layers that lead us to the water. The prediction of soil color is important to reduce the risks involved in the drilling process. The DE-based optimization algorithm is used to select the best candidate and assign weights to each classifier based on their performance. The voting strategy merges the knowledge of all the member classifiers and makes one ensemble model. The ensemble model assumes to be better than the individual classifiers. The soil color has ten different values. This is because we have multiple colors of soil at different depths. The hardness and softness of the soil color are also essential for the cost estimation of the whole process before starting drilling. Like soil color, the land layer is another attribute representing the underground layers of the land at different depths. We have eight different land layers in the data, including landfill layer, Sedimentary layer, Weathered soil layer, Weathered rock layer, Soft rock layer, Gyeongam Formation, Ordinary rock formation, and Burlap soil layer. All these layers have a different level of hardness and average digging depth. The hardness and softness of the soil color and land layer are computed by using depth and the number of days spent to achieve that depth.

For the prediction of water level and total depth of the borehole, multi-regression models are used. The multi-regression models predict two values at the same time based on independent input variables. In the chaining mechanism, two trained models are combined to predict two targets simultaneously. The results of the proposed chained multi-out regression model based on Support Vector Regressor (SVR) are compared with Decision Tree Regressor (DT-R), K-Nearest Neighbor Regressor (KNN-R), and Linear Regressor (LR). The results are compared in terms of Mean Absolute Error (MAE), Mean Square Error (MSE), and R2 score.

The core contributions of the proposed study are listed below.

- Different features are extracted from the dataset, and the hardness level of the underground layers is computed
- The study proposed weighted voting classifier based on Differential Evaluation optimization algorithm
- The weights of the candidate classifiers are learned from DE based on the performance of the classifier
- The chained multi-objective regression model is presented to predict the water level and drilling depth in different regions
- The study facilitates the drilling industry to select the region with a soft layer and acceptable range of water level and depth

The rest of the paper is structured as follows: section II discuss the models and techniques used to facilitate the drilling industry; Section III describe the proposed weighted voting classifier for classification and chained multi-output model for the prediction of water level and drilling depth. Section IV presents the results of proposed models and compares them with the existing state of the are ML techniques to show the effectiveness of the proposed model. Finally section V conclude the contributions with some possible future directions.

II. RELATED WORK

Groundwater is complex at the same time, a fragile resource that is substantial to domestic, economic, agriculture, and industrial activities [12]. Moreover, it is a vital source of replenishing safe drinking water worldwide and plays a critical role in natural ecosystems. There are various ways through which groundwater exploration can be done. One of the widely used groundwater exploration and development techniques is drilling. However, drilling is not confined to just boring a hole in the ground and acquiring water from it [13]. Drilling groundwater is becoming a multi-billion industry due to the growing demand for groundwater acquisition through downhole drilling operations. The surge for fulfilling massive water demand and providing safe drinking water has aggravated the drilling operations for groundwater resource exploitation. However, groundwater acquisition is not as simple as it looks [14]. Before drilling, many preliminary surveys and analyses need to be conducted so that desired information related to the estimation of groundwater characteristics must be estimated beforehand. A particular geological setting, hydraulic continuity, groundwater depth, quality and quantity along with geological layer analysis are substantial part of groundwater characteristics.

The drilling process for groundwater acquisition can simply be defined in four major steps. Site visiting by a hydrogeological expert for assessment of geological characteristics so that the risk of drilling into natural hazards

must be avoided [15]. The second phase involves drilling and construction of boreholes, followed by aquifer testing to determine water borehole yield. Lastly, the hydrogeologists ascertain the pumping and piping system types based on the intended use of water resources. However, the drilling process is highly resources intensive and consumes a huge amount of budget. The cost of drilling is influenced by the type of ground, groundwater, and borehole depth, the machinery involved, experts, human resources, and materials required [16]. Due to advanced drilling and pumping techniques, the reliance on groundwater has increased manifolds. Increased drilling operations all over the world are producing a massive amount of data. Hence, data-driven modeling and the application of advanced data analytics tools to predict downhole environmental aspects are deemed necessary. Geological and groundwater bore complexities are incurring additional costs and wastages of scarce resources. Consequently, optimizing the drilling operations through the application of machine learning techniques is essential to reduce resource wastage and increase drilling productivity [17]. Prediction and classification of various hydrogeological factors can assist the drilling companies to avoid issues like stuck pipe, over cost, and low water levels during the drilling process in different areas.

Recently, many techniques have been applied to optimize the drilling process for the achievement of optimal drilling parameters such as rate of penetration [18]–[20]. Predictive modeling and big data analytics for time series drilling data have driven huge interest by the scientific community. The Researchers have successfully implemented machine learning and artificial intelligence methods with a major focus on reducing various parameters that are substantial for groundwater drilling such as the non-productive and invisible lost time during drilling [21]. For efficient groundwater acquisition, there is a need for an optimized drilling process such as rate of penetration. Article [22], [23] proposed a machine learning model for predicting the rate of penetration based on data analytics using a real-time data set. The authors consider seven input parameters for the proposed work including surface parameters and others, including torque, stand pipe pressure, differential pressure etc. Experimental results reveal that the rate of penetration increased by 14%. In the article, [24] rate of penetration is predicted using statistical regression and artificial neural networks. To find the correlation between various variables, the authors performed exploratory data analysis. Furthermore, the importance of predictors is also computed. For prediction, several models are employed, such as neural network, step-wise regression, classification and regression techniques (CART), and K-nearest neighbors. The findings of the proposed work suggest that ensemble methods can improve the performance of the system. Another study [25] proposed a probabilistic frequency-based ratio model for mapping of groundwater potential using eight factors related to topography, geology, and satellite imagery. Moreover, the authors analyzed several relationships between the yield of groundwater and hydrogeological factors. Results of the frequency model depicted that the area under the curve is about 84.78 percent.

Article [26] proposed a random forest-based mapping scheme for groundwater yield. Experimental findings suggested that the generalization performance achieved by self-learning random forest achieved 23 percent improvement. Furthermore, the comparative analysis of the proposed approach with random forest, support vector machine, artificial neural network, decision tree, and voted artificial neural network and random forest verified the effectiveness of the proposed solution. Article [27] developed a prediction framework based on machine learning algorithms for predicting the water depth. The input variables considered for the study are meteorological data, historical and upstream water level gauge data from 2009 to 2015. Artificial Neural Network, Random Forest, Decision Tree, and Support Vector Machine are trained to predict the water level. Experimental findings suggest that Random forest achieved root mean square error RMSE of 0.09 percent. The study [28] proposed a deep learning framework for predicting the water quality and depth. The deep learning framework comprises CNN deployed for water level simulation and LSTM for water level prediction. The results conclude the effectiveness of the proposed solution for accurately simulating water quality and quantity.

Article [29] proposed a modeling approach for identifying changes in levels of groundwater using machine learning models. The study applied an ensemble modeling scheme using spectral analysis. Experimental results proved that ensemble learning can be used as an alternative approach for the simulation of groundwater changes and can better ascertain water availability in regions with subsurface properties. The ROP is treated as a prediction problem and solved by using different optimization techniques [30]. ANN is used to predict the ROP by optimizing the parameters. The method was applied on vertical and horizontal drilling for oil wells.

Furthermore, the authors established the relation between irrigation demand and its impact on change in groundwater level. Another study [31] presented a comparative analysis of machine learning models for mapping groundwater potential through prediction. The models involved in the research are mixture discriminant analysis (MDA), random forest (RF), multivariate adaptive regression spline (MARS), and boosted regression tree. The study analyzed the spatial distribution of various hydrogeological and physiographical factors along with their respective spatial distribution. Results verify the effectiveness of the MDA model for mapping groundwater potential. Groundwater analysis has been conducted in article [32] related to prediction approaches using machine learning methods. The findings of the proposed study suggest that data-driven modeling approaches are highly effective in predictive modeling and decision making; however, there is a need to improve the accuracy of these systems. Article [33] Proposed SVM-based model to detect anomalies in groundwater using real-time data. Another study [34] presented a machine learning-based solution for predicting lithology formation based on drilling parameters. Results show excellent

abilities of the ANN model for the identification of formation lithology. Article [35] employed RF to predict concentrations of nitrate groundwater. The findings of the proposed study suggest the use of spatial predictors for predicting nitrate concentrations.

Although, multiple ML techniques are applied on different application areas [36], [37] and the enhanced versions are proposed [38] and many researchers concerned with hybrid or ensemble techniques [8]. Majority voting is one of the popular and widely used ensemble techniques. It is just a decision rule to decide about any sample by getting the knowledge of all candidate classifiers. In case of simple voting classifier, the model does not required a parameter tuning [39], [40]. It is a crucial issue to select the best classifier and assign weights based on its performance for a particular class [41]. The weight assignment can be seen as an optimization problem and can be optimized using different optimization algorithms like Genetic Algorithm, Particle Swarm Optimization (PSO), and DE.

III. METHOGOLOGY

The dataset is provided by Jeju National University, Republic of Korea. The data contains multiple attributes related to several drilling points. The dataset contains the information which reflects the condition of the borehole. The total depth of the drilling point, including location coordinates and the groundwater level, is given in the dataset. Different number of days required to complete drilling, so the data contains n number of rows for one drilling point where n is the number of days spent. Each sample represent the starting and ending depth for each day and the soil color and land layer found on that day is also given in the dataset. Further attributes are listed in the Table 1.

First, The data is preprocessed and some features are extracted from the existing one. In the raw dataset, each drilling point contains multiple records, where each record presents the information related to a single drilling point. Figure 1 shows the information of one drilling point with six records. There are multiple land layers and soil colors found during the process of drilling. Similarly, Multiple days spent on each drilling point and different level of depth are achieved as illustrated in Figure 1.

A. FEATURE PREPROCESSING AND FEATURE EXTRACTION

To perform the mathematical models on the extracted samples, data must be in the form of numbers. To achieve this goal, an Ordinal encoder is used to convert the categorical values into numbers. The ordinal encoder assigns a unique number to each category of the feature [42]. For instance, we have ten different soil colors, so the ordinal encoder assigns numbers 1-10 for each soil color. Some features from the dataset are dropped because they show no pattern as per the target attribute. For example, borehole code and drilling resonance have unique values for each sample, so these attributes are removed from the dataset. Because, the unique attributes have no pattern in terms of target feature.

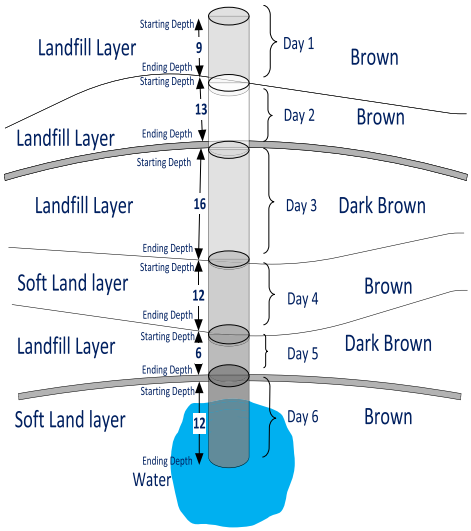


FIGURE 1. Sequence of one drilling point.

Next, some features from the data are extracted to improve the performance of the model. Existing features like staring depth, ending depth are used to extract new features from the data.

The first feature extracted from the data is the total depth of each drilling point. The total depth is the depth in which the water is found. That total depth of drilling point is achieved in multiple days. The eq. (1) is used to extract the total depth using starting and ending depth.

$$TD = \sum_{i=1}^{n} E_{Depth} - S_{Depth} \tag{1}$$

where E_{Depth} is ending depth, S_{Depth} is starting depth, and n is the number of instances of a single borehole. On each drilling point, different days are spent, to calculate the number of days, instances of each point are calculated, and the following query is used to extract the number of days from the database.

Select count$(*)$ *as 'Num of days' from*
'drilling data' group by location;

After the extraction of features, the feature set is completed and passed to the classification model.

The total depth is computed features from the existing one. The total depth is the depth of the borehole on which the water is found. The total depth of borehole along with x and y position is shown in 3D space in Figure 2.

B. CLASSIFICATION

1) TRADITIONAL MACHINE LEARNING MODELS

Multiple machine learning techniques are used to classify real-world data into various classes. In the first phase of the experiments, data is passed to the classification model to predict the land layer and soil color. Figure 3 shows the complete classification model to predict the soil color and the land layer. Initially, the raw data is passed to the preprocessing

TABLE 1. Detail feature description.

Attribute	Description
Location of Drilling point	X and Y coordinates of the drilling point
Starting Depth	The Depth on which the drilling process starts on specific day as shown in Figure 1
Ending Depth	The Depth on which the drilling process ends on specific day as shown in Figure 1
Altitude	Drilling point's altitude
Korean Stratup	The startup attribute defined by Korean organization
Starting Date	The day when the specific drilling point is started
Ending Date	The day when the specific drilling point is ended
Ground Water level	The level of the water for that specific drilling point
Land Layer	The layer found during the drilling process (eight distinct values)
Soil Color	The soil color found during the drilling process (ten distinct values)

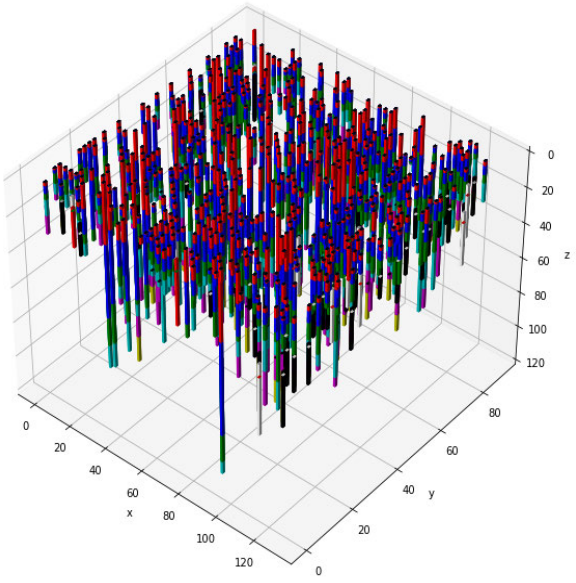


FIGURE 2. 3D visualization of each drilling point.

step and then the feature extraction step, as discussed in the previous step.

The processed data is passed to the classification module along with target classes. Soil color and land layer are used as target labels one by one. From the traditional machine learning models, naïve Bayes with kernel-based estimator [43] is used to estimate the probability of the target class. The general form of the kernel-based estimator is given in eq. (2).

$$f(x, M) = n^{-1} \sum_{i=1}^{n} K_M(x - x^{(i)}) \tag{2}$$

where M is the smoothing matrix, x is the sample from the dataset, and n is the total number of cases from which the estimator is learned. For the supervised learning approach, the data is in the form of $(x^1, y^1) \ldots .(x^n, y^n)$ with n number of samples. The kernel estimator used here is given in eq. (3).

$$\hat{I}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{\hat{f}(x^{(i)}, y^i)}{\hat{f}(x^{(i)})\hat{f}(y^{(i)})} \tag{3}$$

where the function $\hat{f}$ represents the density-based kernel estimator given in eq. (2). The purpose of selecting kernel-based naïve Bayes is to determine better results than the state-of-the-art Naïve Bayes classifier [44].

Support vector classifier is one of the famous classifier due to its powerful kernel tricks. SVM not only focuses on separating line but also set the best hyperplane between the classes. The kernel tricks of the SVM transform the data into kernelized feature space where the data is more likely to be linearly separable. To transform the data from the original to kernelized feature space, quadratic and Gaussian kernels are used. RBF Gaussian kernel is famous and most commonly used to solve multi-class nonlinear problems [45]. Initially, data is transformed from original feature space to kernelized feature space using quadratic kernel eq.(4) and Gaussian kernel eq. (5).

$$K(x, y) = (x^T y + c)^d \tag{4}$$

where x is the input sample, y in target variable, d is the degree polynomial and c is the tradeoff parameter where $c \geq 0$.

Similarly, the Gaussian kernel use gamma to transform the data samples

$$K(x, y) = exp(-\gamma \left\| x_i - x_j \right\|^2) \tag{5}$$

where, the parameter γ is used to scale the mapping, the Gaussian transformed feature space is more linearly separable.

Along with machine learning models, some deep learning techniques are also used to solve complex and multiclass problems. In this study, a Bilayered Neural Network with the Relu activation function is used to predict the soil color and land layers using an input set of independent variables. The architecture of NN used in the experiment is shown in Figure 4. There are two fully connected layers involved in the network, followed by hidden layers. In a fully connected layer, all the input from one layer is connected to the activation unit of the next layer. The output of the Bilayered neural network is optimal weights used to predict the target class in the validation and testing step. Finally, all the selected drilling characteristics are passed to the NN to predict the soil color and land layer.

The purpose of the selection of SVC is that the kernel tricks of SVM transform the data into kernelized feature space where kernelized NB is suitable for the multiclass problem. DT is one of the famous entropy-based classification algorithm and is commonly used in multiclass problems. The working strategy of KNN is based on the neighbor's value.

The soil color and land layer have different patterns and are somehow related, so, the neighbor information is useful. By considering this information, KNN is selected as one of the candidate for voting.

In the first step of the experiments, raw data is passed to the preprocessing phase to process the data before passing for classification. Then, the first step in feature importance is computed, and features are selected to impute the missing value. Instead of removing missing values from the data, we imputed them by the KNN imputation technique.

2) PROPOSED ENSEMBLE WEIGHTED VOTING MODEL

In the ensemble model, the voting classifier is used to classify the data into different classes. Voting itself is not a classifier but a technique to combine the results of multiple classifiers. The wrapper of a voting classifier consists of multiple trained classifiers and assigns the target label to each sample based on the number of voters. The same data samples are passed to all candidate classifiers to train the model and ensemble them to predict the final output. Ensemble voting classifier is best for multi-class problems [46]. We have selected four different classifiers for the voting wrapper: Gaussian Support Vector classifier, K-Nearest Neighbor classifier, Kernel Gaussian NB, and Decision Tree. The voting strategy merges the knowledge of all involved candidates and decides the label for each testing sample. The conceptual model of the voting classifier is shown in Figure 5. But in the traditional voting classifier, each candidate classifier contributes equally in the selection of the target variable. While it is possible in some situations that one or two candidate classifiers are less confident to assign a target to a sample compared to others, in this case, the wrong class can be assigned.

By considering the given limitation of the traditional classifier, we proposed a weighted voting classifier in which we use DE to select the best optimal weights for each classifier to select the best target variable for each sample. The problem is to assign optimal weights to each classifier based on its confidence for each class. The DE is a population-based optimization technique used widely for multiple search problems in the literature. DE first create the list of population of size N and D-Dimensional vector, where $X_{i,G} = [x_{i,G}(1), x_{i,G}(2), \ldots x_{i,G}(j) \ldots x_{i,G}(D)]$. DE performs three main steps, including mutation, crossover, and selection.

Mutation: mutation is the first step of the DE optimization algorithm which generates the donor vector represented by $V_{i,G}$ for each target vector $X_{i,G}$ for the current generation.

Crossover: Crossover simply generate the trail vector $T_{i,G} = [t_{i,G}(1), t_{i,G}(2), \ldots t_{i,G}(j) \ldots t_{i,G}(D)]$ by performing the crossover operation between target and corresponding vector. For each variable j from the D-dimensional vector:

$$\begin{Bmatrix} v_{i,G}(j), & \text{if random value (i,j)[0,1]} < \text{CR} \\ x_{i,G}(j), & otherwise \end{Bmatrix} \quad (6)$$

where CR represents the crossover rate in the range of [0,1], random number is a uniform distribution for each *jth* value. The random value ensures that the trail vector $T_{i,G}$ get at least one value from mutant vector $V_{i,G}$.

Selection: The selection process selects and compares the trails with target individuals and decides based on the survival of the target. If the target is able to survive in the next generation, the trail individual will be selected. The operation of selection can be presented as

$$\begin{Bmatrix} T_{i,G}(j), & \text{if Trail } f(T_{i,G}) \leqslant \text{individual } f(T_{i,G}) \\ X_{i,G}(j), & otherwise \end{Bmatrix} \quad (7)$$

where the function $f()$ is the function to be optimized and ensure that the selected member is best for the individual. The described technique is used in the selection of optimal weights for classifiers; however, multiple crossover and mutation techniques for DE have been proposed [47]–[50]. The Algorithm 1 shows the flow of the selection of best weights using DE from the given search space. The DE algorithm returns the best weight vector $(w_1, \ldots w_n)$ for each member classifier.

Algorithm 1 DE Based Proposed Weighted Voting Classifier

Input: DE Control Parameters: *Population size N*, *Mutation factor F* and *Crossover Rate CR*
Output: Optimal weights $(w_1, \ldots, w_n)$;
initialization;
Population of N individuals;
$X_G = X_{1,G}, X_{2,G}, \ldots, X_{N,G}$
// Uniformly Distributed
where $X_{i,G} = [x_{i,G}(1), x_{i,G}(2), \ldots, x_{i,G}(N)]$ represents the weights $(w_1, \ldots, w_n)$ of classifiers.
G $\leftarrow$ 0 // generation iteration
while *stopping criteria not satisfied* **do**
  for $i = 0$; $i < N$; i++ **do**
    Select distinct indexers $r1, r2, r3$ /* should be different from i */
    $V_{i,G} = X_{r1,G} + F \times (X_{r2,G} - X_{r3,G})$
    // Compute mutant vector
    $j_{rand} \leftarrow random()$ // Random value
    for $j = 0$; $j < D$; $j++$ **do**
      $T_{i,G} \leftarrow$ using eq. (6)
    $X_{i,G+1} \leftarrow$ Fitness of $T_{i,G}$ and $T_{X,G}$ using eq. (7)
  G $\leftarrow$ G+1
  // Increase termination parameter

The fitness evaluation of the model is based on the accuracy of the member classifiers. For c number of categories in the target attribute with D number of classifiers to vote, the predicted value V_p of the proposed weighted voting for k sample is

$$V_p = arg\ max \sum_{i=1}^{D} (\vartheta_{ij} \times w_i) \quad (8)$$

where ϑ_{ij} is the binary decision variable; $\vartheta_{ij} = 1$ if j^{th} category is assigned to the k sample by i^{th} classifier and

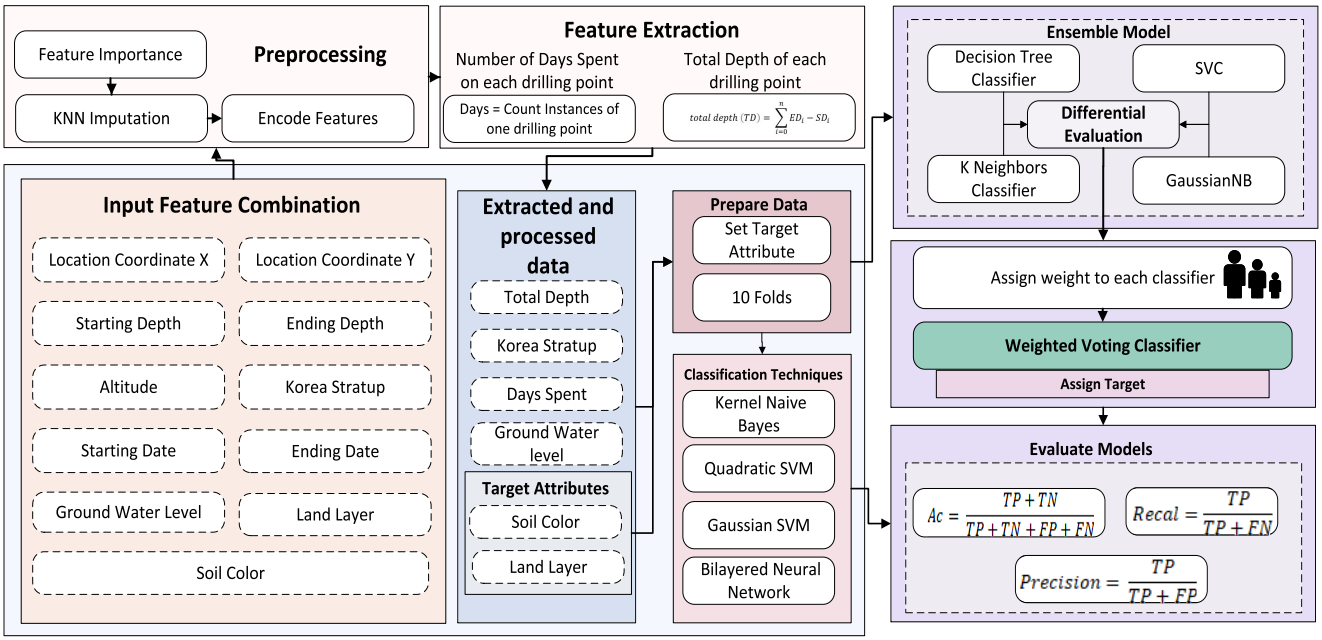


FIGURE 3. Data preprocessing, feature extraction, state of the art ML techniques and operational flow of enhanced optimization-based voting classifier.

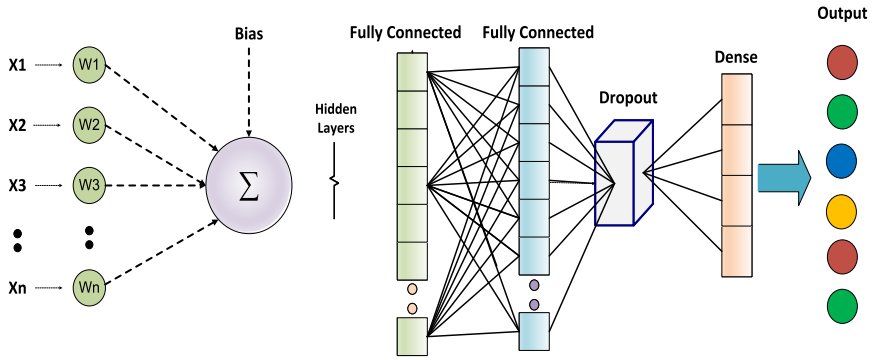


FIGURE 4. Architecture of Bilayered NN.

$\vartheta_{ij} = 0$ otherwise. w_i represents the weight for i^{th} classifier and optimized by the Algorithm 1.

C. REGRESSION

1) REGRESSIONS MODELS

Regression analysis is a type of prediction in which the relationship between dependent and independent variables is investigated. Moreover, these techniques are used in time series prediction, forecasting, and determining the variable's causal effect relationship. Multiple regression techniques are available to determine or predict the next value of a dependent variable. These models fit the curve based on given data. Linear regression, K-nearest Neighbours Regression Decision tree regressor, and Support vector Regressor are famous and most used Regression models. We use the traditional regression models to predict water level and drilling depth and compare the results of the conventional models with the proposed SVR-based chained multi-objective regression model in terms of MAE, MSE, and R2 score.

In the proposed chained multi-out regression model Support vector regressor is selected to make a chain of the models for each target attribute. The single model is able to predict both target attributes at one time, as shown in Figure 6. As the SVR shows better results as compared to other ML model, so the best model is selected as candidate for proposed technique. In the multi-out regression model, the SVR model is first trained based on one target attribute, and then the prediction of the first model along with training data is passed to the second model to predict another attribute. The chained model automatically splits the target attribute and trains both models by chaining both trained models and simultaneously predict multiple variables.

The prediction of the water level and depth is starts from the separation of the target variables from independent features and shown in Figure 7. After the separation, the stand-alone regression models are applied to predict the water level and total depth separately. After that SVR based multi objective chained model is applied to predict both variables

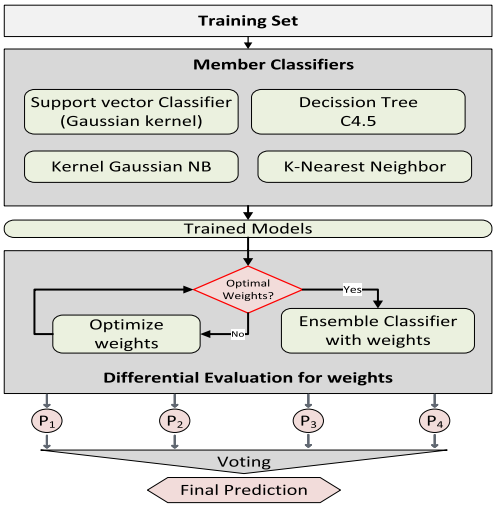


FIGURE 5. Structure of proposed DE-based weighted voting classifier.

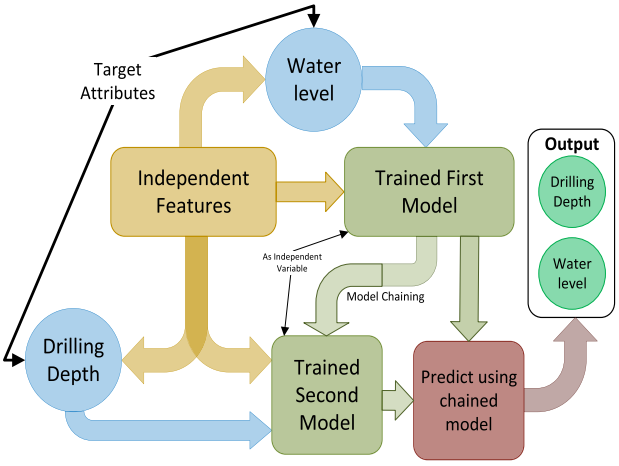


FIGURE 6. Multiout Regression model based on SVR.

simultaneously. K-Fold cross-validation is used to train the model. Finally, existing and proposed models are evaluated using MAE, MSE, and R2 score.

IV. RESULTS

A. EVALUATION METRICS

To evaluate the performance of the proposed classification model and for a fair comparison, accuracy, precision, and recall are used as evaluation metrics. These metrics are widely used for the evaluation of classification models [51]. The accuracy metric refers to how much the measurements are close to the true or accepted value eq. (9).

$$Ac = \frac{TP + TN}{N} \tag{9}$$

where TP is true positive, TN is true negative and N is total number of samples. Accuracy is easier to understand that how positively the model is performing on the data. To check the positive class prediction from all positive samples, recall is

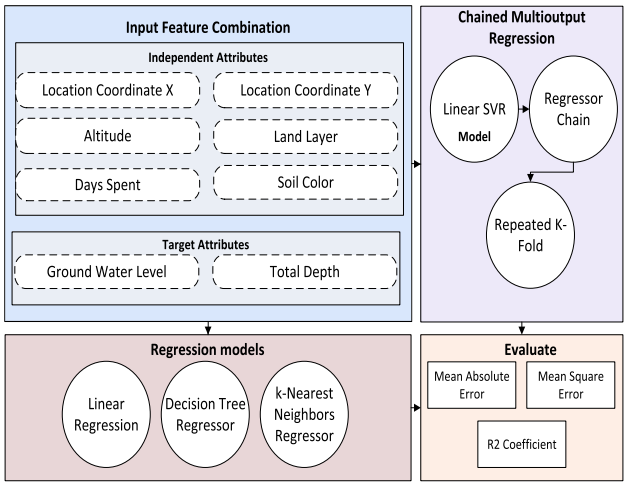


FIGURE 7. Multiout regression model to predict depth and water level simultaneously.

used eq. (10).

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where TP refers to True positive and FN is false-negative rate.

Precision is a little bit change to recall; it reflects the information related to the true positive class. The prediction of true positive samples from all true positive and false-positive samples eq. (11).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

As we have multiple classes of soil color and land layer, so the prediction value is based on the mean of all classes.

To evaluate the performance of the chained multi-out regression model, MAE, MSE, and R2 are used. MAE shows the absolute difference between actual and predicted water level and drilling depth. The eq. (12) shows how the MAE is computed from the actual and predicted value.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{12}$$

The mean squared deviation (MSD) or mean square error (MSE) squares the error between actual and predicted values.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \tag{13}$$

The R2 coefficient i different from all other error metrics, instead of error it shows the accuracy of the model. the higher value of R2 shows better performance of the model. The eq. (14)

$$R2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation} \tag{14}$$

TABLE 2. Average accuracy, precision, and recall of validation and test phase for soil color.

	Validation			Test		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Kernel Naive Bayes	67.21±0.83	64.21±1.31	66.24±1.15	69.91±0.81	64.61±1.25	63.74±0.83
Fine Gaussian SVM	68.37±0.86	67.87±1.08	70.24±1.19	71.42±0.89	70.49±1.01	69.81±1.17
Quadratic SVM	70.29±0.84	68.11±1.07	67.88±1.07	70.09±0.79	70.14±1.16	67.61±0.83
Bilayered Neural Network	75.29±1.03	72.73±1.41	77.82±2.04	80.14±1.27	78.09±0.84	80.94±0.74
Voting Classifier	77.19±1.09	75.41±1.23	79.21±0.89	81.27±1.31	81.94±0.18	79.58±0.76
Proposed Weighted Voting Classifier	79.24±1.28	80.47±0.73	79.29±1.27	83.41±0.89	82.85±0.59	83.11±1.18

B. CLASSIFICATION RESULTS

To predict the underground soil layer and land layer in different regions, multiple classification algorithms are applied, and the results are listed in Table 2 and 3. The drilling task becomes easy when the land layer and soil color are known in advance because soil color and land layer have the different softness and hardness levels. The average digging capacity per day of each layer and soil color shows the hardness level. Figure 8 shows the average digging capacity of each soil color and land layer.

There are eight different land layers in the given dataset. Four different state-of-the-art ML algorithms are selected to predict land layers on different locations and depths to solve this multi-class problem. The performance of Bilayered NN is better as compared to all other ML algorithms in terms of precision, recall, and accuracy. The ensemble voting classifier is also applied to compare the results with the proposed optimization-based weighted voting classifier. The same member classifiers are used to for both voting classifiers. The results show that the proposed weighted voting classifier shows better classification results as compared to the existing voting classifier and traditional ML classification algorithms.

The results of the prediction of soil color are illustrated in Table 2. The results show the effectiveness of the proposed weighted voting classifier. The accuracy achieved by the proposed model for validation is 82.76% and 86.69% for the test phase. The proposed model use the knowledge of all classifier based on their performance for each class so the results are better than all other state of the art techniques. The early prediction of the land with soft soil will allow the drilling industry to figure out the resources before starting the actual process. Along with it, the rate of penetration can also be increased by considering the predicted information. There are multiple risks involved in drilling, including stuck pipe, fluid broken, and over cost. All these risks are involved because of the hard layer. Because of the hard layer, pipes are stuck, and pipes fluid can also be broken. The early prediction can minimize the risks involved in the drilling process. If the predicted layer is hard, the drilling industry should use the related material to reduce the time and cost.

The proposed weighted voting classifier merges the knowledge of all the involved candidate models based on their confidence level of prediction and decides the label. For better measurement, 10-fold is used to train candidate models, and the average accuracy is used as a final parameter for the optimization module. For each iteration, nine out of ten folds are used as training, and the process continues for ten-time so that each fold is used in a testing phase. The standard deviation of upper and lower bound of precision, recall, and accuracy achieved in the folds for each classifier is reported. For the effectiveness of the proposed methods, the model is compared with existing ML techniques for both soil color and land layer as shown in Table 2 and 3 respectively. The validation and test accuracy, precision, and recall are reported and compared. As the proposed model considers the candidate classifier's performance so that the optimal weights are assigned to each classifier. As result, the role of the candidate with higher accuracy in the selection of target variable is more as compared to other algorithms.

Average per day digging capacity of each soil color and land layer is computed to analyze the hardness level. Per day digging capacity of both soil color and land layer is computed by using eq. (15).

$$DC(color, layer) = \frac{\sum_{i=1}^{n}(ED_i - SD_i)}{n} \tag{15}$$

where DC is Digging Capacity, ED is ending depth, SD is starting depth, and n is the number of instances of that specific soil color or land layer. The Figure 8 illustrates the information computed by the eq. (15) for both soil color and land layer. The figure shows that the dark brown soil color is soft because the average depth achieved per day is 4.3 meters. In contrast, the depth achieved in the case of partridge soil color is 2.3, which shows that partridge soil is hard and consumes more days and resources. The drilling industry needs attention when the partridge soil color is predicted because the fluid of the pipe can be broken because of its hardness.

In the case of the land layer, the Gyeongam Formation layer is soft compared to all others. While the landfill layer is too hard because we can achieve only 1.5-meter depth by working a whole day. By analyzing the results of the land layer, we can say that the rate of penetration in case of the landfill layer is low. Figure 2 shows the drilling point with a different layered structure. Each slice of 3D pipe shows the digging depth on a specific day. At the same time, the total number of slices in the 3D pipe shows the number of days spent in that particular borehole. The z-axis of the figure shows the total depth of drilling points. Different colors on one borehole point show the number of days spent on each drilling point and the thickness of each layer. The situation of the underground water table is also depicted in the illustration. In some areas, the water level is low, while the water level is too high in some areas.

TABLE 3. Average accuracy, precision, and recall of validation and test phase for land layer.

	Validation			Test		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Kernel Naive Bayes	71.54±0.91	70.12±1.04	72.10±0.74	76.34±0.79	71.81±1.32	75.91±1.14
Fine Gaussian SVM	69.77±0.74	69.51±0.87	68.09±1.08	74.39±1.42	73.18±0.49	74.59±0.81
Quadratic SVM	72.54±1.24	69.76±0.87	74.42±0.67	77.83±1.20	73.08±0.73	79.66±1.81
Bilayered Neural Network	76.31±0.49	76.97±0.89	75.83±1.06	81.16±0.79	79.32±0.91	80.98±1.13
Voting Classifier	80.47±1.54	81.69±1.09	80.04±1.42	83.26±0.96	84.39±1.27	82.94±1.29
Proposed Weighted Voting Classifier	82.76±0.97	82.97±1.08	81.78±1.47	86.69±1.21	87.29±0.27	86.27±0.72

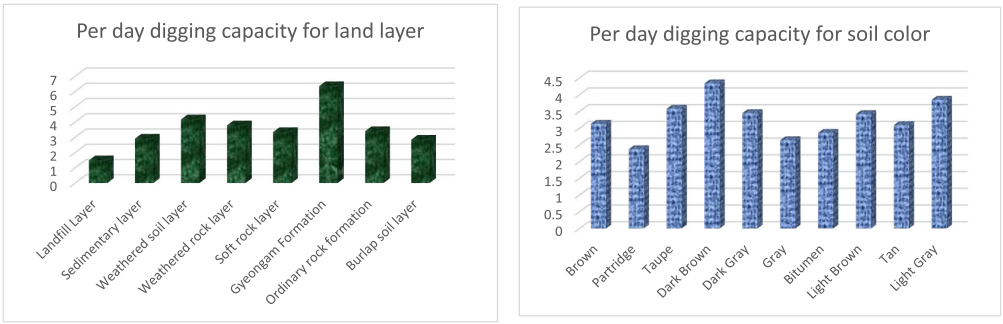
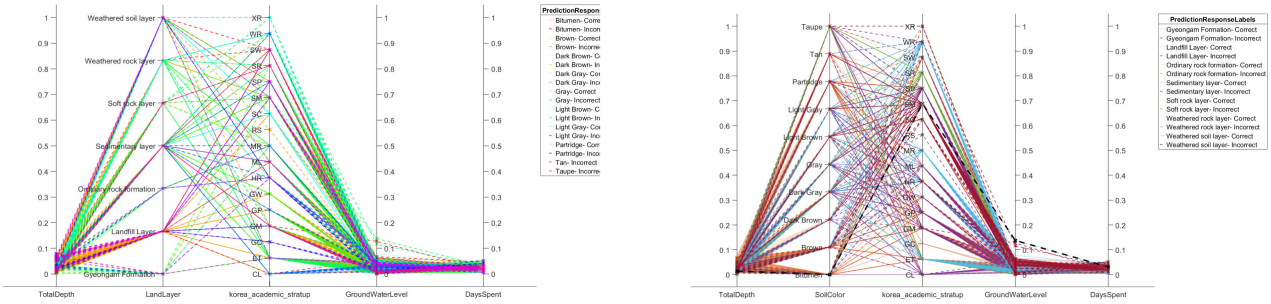


FIGURE 8. Average digging capacity of both soil color and land layer.



(a) Sequence of Trained model on soil color

(b) Sequence of trained model on land layer

FIGURE 9. The sequence of trained model.

To sum up the knowledge of all the candidate classifiers, the patterns illustrated in Figure 9a and 9b are extracted. The sequence of prediction or the flow of the trained model is visualized. The sequence shows a separate line for each class label. Numeric values are normalized, and the categorical value remains the same. The figure shows some patterns between independent features. When the value of Korea startup has changed, the value of the layer also changes in most cases. The color is assigned to each correct and incorrect prediction, as shown in the right bar of the figures.

C. REGRESSION RESULTS

The prediction of water level and depth is based on region, Altitude, Land layer, days spent, and soil color. The borehole depth is highly dependent on days spent on the process and the hardness level of both soil color and land layer. Similarly, different land layers and soil color patterns lead us to different groundwater levels. The results of prediction of water level and drilling depth shown in Table 4 and 5 respectively. The results of the proposed chained multi-out regression model are compared with conventional models of ML. The depth of the borehole varies at different locations as per the level of groundwater. Similarly, the pattern of soil color and land layer also plays an important role in the prediction of depth.

The results show the effectiveness of the proposed chained model in terms of MAE, MSE, and R2. As the proposed method consists of two trained regression models where the sequence is sequential. Because of the sequential method, the previously trained model injects the knowledge to the next model to predict one dependent variable. The next model uses that predicted variable as an independent variable to train itself to predict another variable. Because of that chained strategy, the proposed method's performance is good compared to other conventional ML techniques. Figure 10 shows the prediction results for the training and testing phase for drilling depth, and Figure 11 shows the prediction results of water level.

The proposed model is based on two chained models to predict two different target variables simultaneously. By predicting the level of water and depth in different regions,

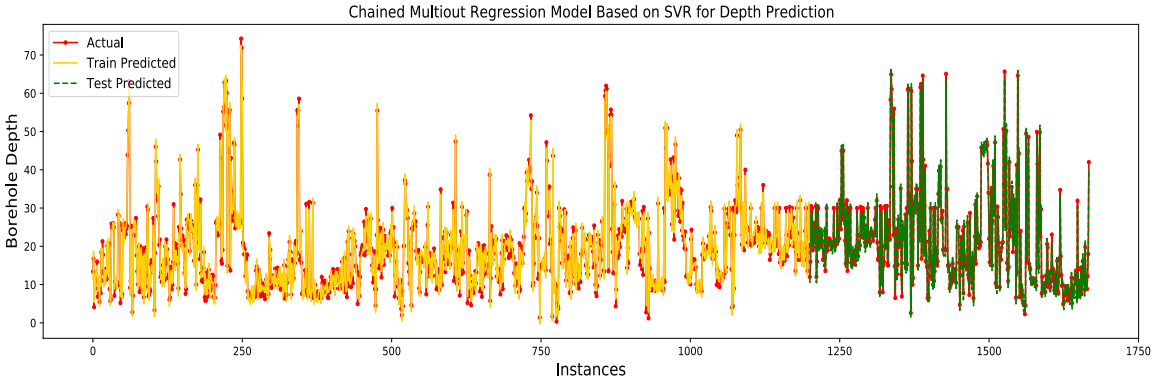


FIGURE 10. Results of multiout regression model for the prediction of next depth.

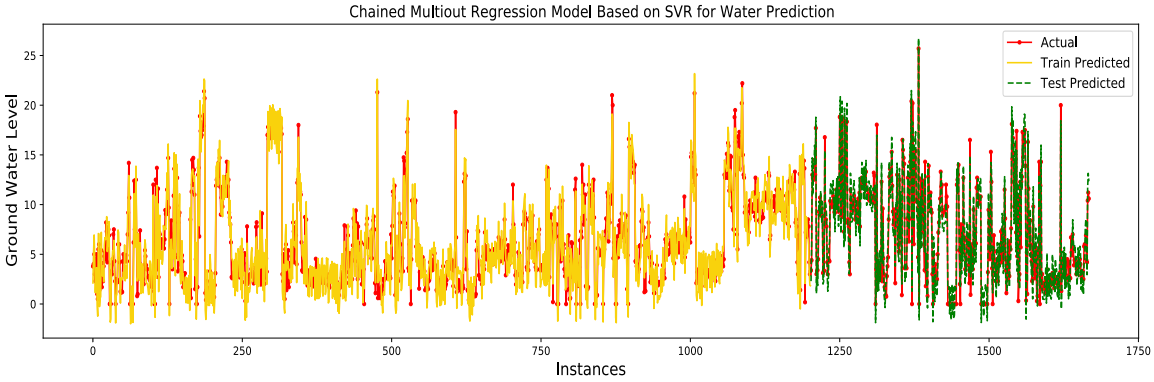


FIGURE 11. Results of multiout regression model for the prediction of water level.

TABLE 4. Comparison of proposed chained multiout regressor and traditional ML regressor in terms of MAE, MSE, and R2 for groundwater level prediction.

Regressor	MAE	MSE	R2
Decision Tree Regressor	7.254	103.76	67.19
K-Nearest Neighbors Regressor	4.810	61.85	78.03
Linear Regression	8.409	124.86	61.98
Proposed Multi-Objective Model	3.893	54.816	84.86

TABLE 5. Comparison of proposed chained multiout regressor and traditional ML regressor in terms of MAE, MSE, and R2 for borehole depth prediction.

Regressor	MAE	MSE	R2
Decision Tree Regressor	6.637	83.03	72.45
K-Nearest Neighbors Regressor	4.203	56.64	81.61
Linear Regression	7.933	108.02	64.10
Proposed Multi-Objective Model	3.264	51.746	89.32

the early estimation can be done by the drilling industry. Moreover, the prediction of soil color and land layer and their hardness level facilitates the drilling industry to manage the required resources. The selection of the area depends on the preferences of the drilling industry. If the industry has heavy resources, then the hard layer can be compromised, and areas with low depth can be considered. While, on the other hand, if the labor cost is not a problem and the drilling equipment are not as much strong to drill the hand layer, then the area with soft layer should be selected. Therefore, the classification and regression results help the industry. While the selection of the region is highly dependent on the preference of the drilling industry and their method of drilling.

V. CONCLUSION

Huge data is generated by scientific experiments and other mediums. Researchers have applied multiple clustering and classification techniques to extract the patterns from the data to make it useful. Different ML techniques are also enhanced to get more accurate results from the raw data. The proposed work presents an optimization-based weighted voting approach to find the hidden patterns from the data. The data is accrued from JNU, Republic of Korea. In the first phase, data is preprocessed, and some features are extracted, like total depth and number of days spent. Moreover, the data is encoded using an ordinal encoder to apply mathematical classification algorithms. For fair voting, a population-based optimization algorithm is used to find the optimal weights based on the performance of the member classifiers. The weight is assigned to each classifier, and the target variable is assigned to each test data sample. The proposed method is evaluated based on accuracy, precision, and recall. It is noticed from the experiments that the proposed model outperforms as compared to the existing voting classifier and state-of-the-art ML algorithms. The prediction of soil color and land layer on different areas and predicted depth allows the drilling industry to estimate the time and labor cost early. However, some factors should be considered before starting

the actual drilling process, together with soil color, land layer, water level, and depth of the drilling point. The study uses a weighted voting classifier to predict the soil color and land layer on different locations. At the same time, the hardness and softness of the underground layer are computed to select the related equipment for drilling. The chained multi-out regression model based on SVR is also proposed to predict water level and drilling depth on various locations simultaneously. Two trained models are chained so that the first model injects his knowledge into the second model to make the predictions more accurate. The proposed model is compared with conventional ML regressors, including DT-R, KNN-R, and LR. The results are compared in terms of MAE, MSE, and R2 score. The results show the significance and effectiveness of the chained regression model.

ACKNOWLEDGMENT

Any correspondence related to this paper should be addressed to Do Hyeun Kim.

REFERENCES

[1] M. Z. Lukawski, B. J. Anderson, C. Augustine, L. E. Capuano, K. F. Beckers, B. Livesay, and J. W. Tester, "Cost analysis of oil, gas, and geothermal well drilling," *J. Petroleum Sci. Eng.*, vol. 118, pp. 1–14, Jun. 2014.

[2] W. F. Prassl, J. M. Peden, and K. W. Wong, "A process-knowledge management approach for assessment and mitigation of drilling risks," *J. Petroleum Sci. Eng.*, vol. 49, nos. 3–4, pp. 142–161, Dec. 2005.

[3] S. Paul Singh and P. Xavier, "Causes, impact and control of overbreak in underground excavations," *Tunnelling Underground Space Technol.*, vol. 20, no. 1, pp. 63–71, Jan. 2005.

[4] L.-H. Luu, P. Philippe, G. Noury, J. Perrin, and O. Brivois, "Erosion of cohesive soil layers above underground conduits," *EPJ Web Conf.*, vol. 140, Oct. 2017, Art. no. 09038.

[5] D. Ruta and G. Gabrys, "Classifier selection for majority voting," *Inf. Fusion*, vol. 6, pp. 63–81, Mar. 2005.

[6] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, 2010.

[7] T. Dietterich, "Ensemble learning," in *The Handbook of Brain Theory and Neural Networks*, vol. 2, no. 1. Cambridge, MA, USA: MIT Press, 2002, pp. 110–125.

[8] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, no. 1, pp. 239–263, 2002.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[10] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[11] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[12] D. Perrone and S. Jasechko, "Deeper well drilling an unsustainable stopgap to groundwater depletion," *Nature Sustainability*, vol. 2, no. 8, pp. 773–782, Aug. 2019.

[13] B. Shirmohammadi, M. Vafakhah, V. Moosavi, and A. Moghaddamnia, "Application of several data-driven techniques for predicting groundwater level," *Water Resour. Manage.*, vol. 27, no. 2, pp. 419–432, Jan. 2013.

[14] R. Ratolojanahary, R. Houé Ngouna, K. Medjaher, F. Dauriac, and M. Sebilo, "Groundwater quality assessment combining supervised and unsupervised methods," *IFAC-Papers Line*, vol. 52, no. 10, pp. 340–345, Jan. 2019.

[15] A. R. Bisson and H. J. Lehr, *Modern Groundwater Exploration: Discovering New Water Resources in Consolidated Rocks Using Innovative Hydrogeologic Concepts, Exploration, Drilling, Aquifer Testing and Management Method*. Hoboken, NJ, USA: Wiley, 2004.

[16] P. Madhnure, "Groundwater exploration and drilling problems encountered in basaltic and granitic terrain of Nanded district, Maharashtra," *J. Geol. Soc. India*, vol. 84, no. 3, pp. 341–351, Sep. 2014.

[17] L. F. F. M. Barbosa, A. Nascimento, M. H. Mathias, and J. A. de Carvalho, "Machine learning methods applied to drilling rate of penetration prediction and optimization—A review," *J. Petroleum Sci. Eng.*, vol. 183, Dec. 2019, Art. no. 106332.

[18] R. Ashena, M. Rabiei, V. Rasouli, A. H. Mohammadi, and S. Mishani, "Drilling parameters optimization using an innovative artificial intelligence model," *J. Energy Resour. Technol.*, vol. 143, no. 5, May 2021, Art. no. 052110.

[19] N. Iqbal, A. Rizwan, A. N. Khan, R. Ahmad, B. Kim, K. Kim, and D. Kim, "Boreholes data analysis architecture based on clustering and prediction models for enhancing underground safety verification," *IEEE Access*, vol. 9, pp. 78428–78451, 2021.

[20] A. Rizwan, N. Iqbal, A. N. Khan, R. Ahmad, and D. H. Kim, "Toward effective pattern recognition based on enhanced weighted K-mean clustering algorithm for groundwater resource planning in point cloud," *IEEE Access*, vol. 9, pp. 130154–130169, 2021.

[21] C. I. Noshi and J. J. Schubert, "The role of machine learning in drilling operations: A review," in *Proc. Day Wed*, Oct. 2018.

[22] S. Chandrasekaran and G. S. Kumar, "Drilling efficiency improvement and rate of penetration optimization by machine learning and data analytics," *Int. J. Math., Eng. Manage. Sci.*, vol. 5, no. 3, pp. 381–394, Jun. 2020.

[23] N. Iqbal, A. Khan, A. Rizwan, R. Ahmad, B. Kim, K. Kim, and D. Kim, "Groundwater level prediction model using correlation and difference mechanisms based on boreholes data for sustainable hydraulic resource management," *IEEE Access*, vol. 9, pp. 96092–96113, 2021.

[24] B. Mantha and R. Samuel, "ROP optimization using artificial intelligence techniques with statistical regression coupling," in *Proc. Day 3 Wed*, Sep. 2016.

[25] M. A. Manap, H. Nampak, B. Pradhan, S. Lee, W. N. A. Sulaiman, and M. F. Ramli, "Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS," *Arabian J. Geosci.*, vol. 7, no. 2, pp. 711–724, Feb. 2014.

[26] M. I. Sameen, B. Pradhan, and S. Lee, "Self-learning random forests model for mapping groundwater yield in data-scarce areas," *Natural Resour. Res.*, vol. 28, no. 3, pp. 757–775, Jul. 2019.

[27] C. Choi, J. Kim, H. Han, D. Han, and H. S. Kim, "Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea," *Water*, vol. 12, no. 1, p. 93, 2020.

[28] S.-S. Baek, J. Pyo, and J. A. Chun, "Prediction of water level and water quality using a CNN-LSTM combined deep learning approach," *Water*, vol. 12, no. 12, p. 3399, Dec. 2020.

[29] S. Sahoo, T. A. Russo, J. Elliott, and I. Foster, "Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S." *Water Resour. Res.*, vol. 53, no. 5, pp. 3878–3895, May 2017.

[30] A. Al-AbdulJabbar, A. A. Mahmoud, and S. Elkatatny, "Artificial neural network model for real-time prediction of the rate of penetration while horizontally drilling natural gas-bearing sandstone formations," *Arabian J. Geosci.*, vol. 14, no. 2, pp. 1–14, 2021.

[31] A. Al-Fugara, H. R. Pourghasemi, A. R. Al-Shabeeb, M. Habib, R. Al-Adamat, H. Al-Amoush, and L. Adrian Collins, "A comparison of machine learning models for the mapping of groundwater spring potential," *Environ. Earth Sci.*, vol. 79, no. 10, p. 206, May 2020.

[32] K. Kenda, M. Cerin, M. Bogataj, M. Senozetnik, K. Klemen, P. Pergar, C. Laspidou, and D. Mladenic, "Groundwater modeling with machine learning techniques: Ljubljana polje aquifer," *Proceedings*, vol. 2, no. 11, p. 697, 2018.

[33] J. Liu, J. Gu, H. Li, and K. H. Carlson, "Machine learning and transport simulations for groundwater anomaly detection," *J. Comput. Appl. Math.*, vol. 380, Dec. 2020, Art. no. 112982.

[34] A. A. Mahmoud, S. Elkatatny, and A. Al-AbdulJabbar, "Application of machine learning models for real-time prediction of the formation lithology and Tops from the drilling parameters," *J. Petroleum Sci. Eng.*, vol. 203, Aug. 2021, Art. no. 108574.

[35] L. Knoll, L. Breuer, and M. Bach, "Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning," *Sci. Total Environ.*, vol. 668, pp. 1317–1327, Jun. 2019.

[36] A.-N. Khan, N. Iqbal, A. Rizwan, R. Ahmad, and D.-H. Kim, "An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings," *Energies*, vol. 14, no. 11, p. 3020, Jan. 2021.

[37] Z. Ghaffar, A. Alshahrani, M. Fayaz, A. M. Alghamdi, and J. Gwak, "A topical review on machine learning, software defined networking, Internet of Things applications: Research limitations and challenges," *Electronics*, vol. 10, no. 8, p. 880, Jan. 2021.
[38] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, "WR-SVM model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification," *Appl. Sci.*, vol. 11, no. 10, p. 4657, Jan. 2021.
[39] L. I. Kuncheva and J. J. Rodriguez, "A weighted voting framework for classifiers ensembles," *Knowl. Inf. Syst.*, vol. 38, no. 2, pp. 259–275, Feb. 2014.
[40] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 27, no. 5, pp. 553–568, Sep. 1997.
[41] A. Ekbal and S. Saha, "Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach," *ACM Trans. Asian Lang. Inf. Process.*, vol. 10, no. 2, pp. 1–37, Jun. 2011.
[42] C. Friedman, "System and method for language extraction and encoding utilizing the parsing of text data in accordance with domain parameters," U.S. Patent 6 182 029, Jan. 30, 2001.
[43] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protei-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.
[44] A. Pérez and I. N. Inza, "Bayesian classifiers based on kernel density estimation: Flexible classifiers," *Int. J. Approx. Reasoning*, vol. 50, no. 2, pp. 341–362, 2009.
[45] B. Scholkopf, K. K. Sung, and C. Burges, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2758–2765, Nov. 1997.
[46] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. 15, Jul. 2018.
[47] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer, "Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems," *IEEE Trans. Evol. Comput.*, vol. 10, no. 6, pp. 646–657, Dec. 2006.
[48] J. Zhang and A. C. Sanderson, "JADE: Adaptive differential evolution with optional external archive," *IEEE Trans. Evol. Comput.*, vol. 13, no. 5, pp. 945–958, Oct. 2009.
[49] J. Brest and M. S. Maucec, "Self-adaptive differential evolution algorithm using population size reduction and three strategies," *Soft Comput.*, vol. 15, pp. 2157–2174, Dec. 2011.
[50] W. Gong, "Repairing the crossover rate in adaptive differential evolution," *Appl. Soft Comput.*, vol. 15, pp. 149–168, Feb. 2014.
[51] N. Iqbal, R. Ahmad, F. Jamil, and D.-H. Kim, "Hybrid features prediction model of movie quality using multi-machine learning techniques for effective business resource planning," *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 9361–9382, Jan. 2021.

ATIF RIZWAN received the M.S. degree in computer science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan, in 2020. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Jeju National University, Republic of Korea. He has two-year academic experience as a Research Associate with COMSATS University Islamabad. He has good industry experience in software development and testing. His research interests include applied machine learning, data and web mining, optimization of core algorithms, and the IoT-based applications.

ANAM NAWAZ KHAN received the B.S. and M.S. degrees in computer science from COMSATS University Islamabad, Attock Campus, Pakistan, in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Jeju National University, Republic of Korea. Her research interests include machine learning applications in smart environments, analysis of prediction and optimization algorithms, big data, and the IoT-based applications.

NAEEM IQBAL (Member, IEEE) received the M.S. degree in computer science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Jeju National University, Republic of Korea. He has professional experience in the software development industry and in academic as well. He has published more than 20 papers in peer-reviewed international journals and conferences. He is serving as a professional reviewer for various well-reputed journals and conferences. His research interests include AI-based intelligent systems, data science, big data analytics, machine learning, deep learning, analysis of optimization algorithms, the IoT, and blockchain-based secured applications.

RASHID AHMAD received the B.S. degree from the University of Malakand, Pakistan, in 2007, the M.S. degree in computer science from the National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan, in 2009, and the Ph.D. degree in computer engineering from Jeju National University, South Korea, in 2015. Currently, he is working as an Assistant Professor at COMSATS University Islamabad, Attock Campus. His research interests include the application of prediction and optimization algorithms to build IoT-based solutions, machine learning, data mining, and related applications.

DO HYEUN KIM received the B.S. degree in electronics engineering and the M.S. and Ph.D. degrees in information telecommunication from Kyungpook National University, South Korea, in 1988, 1990, and 2000, respectively. He was with the Agency of Defense Development (ADD), from 1990 to 1995. From 2008 to 2009, he was a Visiting Researcher with the Queensland University of Technology, Australia. Since 2004, he has been with Jeju National University, South Korea, where he is currently a Professor with the Department of Computer Engineering. His research interests include sensor networks, M2M/IOT, energy optimization and prediction, intelligent service, and mobile computing.