IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 June 2023, accepted 7 July 2023, date of publication 24 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298225

RESEARCH ARTICLE

Predictive Modeling of Water Table Depth, Drilling Duration, and Soil Layer Classification Using Adaptive Ensemble Learning for Cost-Effective Percussion Water Borehole Drilling

QAZI WAQAS KHAN[1], BONG WAN KIM[2], RASHID AHMED[3,4], ATIF RIZWAN[1], ANAM NAWAZ KHAN[1], KWANGSOO KIM[2], AND DO-HYEUN KIM[1,4]

[1]Department of Computer Engineering, Jeju National University, Jejusi 63243, Republic of Korea
[2]Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea
[3]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[4]Bigdata Research Center, Jeju National University, Jeju 63243, Republic of Korea

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

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), and this work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21DCRU-B158151-02), and this research was supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (2021H1D3A2A02082991). Any correspondence related to this paper should be addressed to Dohyeun Kim.

ABSTRACT Water drilling machines are used to drill boreholes in the ground to extract groundwater. The resources required for water drilling vary from region to region due to underground water table depth and ground soil layer. Water drilling on a hard underground soil layer requires different resources than a soft underground. The proposed study facilitates the drilling industry by selecting the region with a soft land layer and increasing the penetration rate. Furthermore, the number of days and water table depth prediction allows the drilling industry to estimate the depth of the water table and time resources to reach the water table at different locations. The classification techniques classify the region based on the soil land layer. Regression techniques are used for predicting water table depth and number of days. The experiments are performed on a borehole log dataset provided by a research organization. This study used Support Vector Machine, TabNet, and Deep Tabular models to predict the land soil layer and compare the results with our proposed Ensemble Weighted Voting Soil Layer Classifier (EWV-SLC). The performance of the classification model is evaluated using accuracy, Precision, Recall, and F1 Score. The experimental finding shows that the EWV-SLC model performs better in accuracy and F1 score than other machine learning techniques. The performance of the regression model is evaluated using Mean Square Error (MSE), Mean Absolute Error (MAE),Root Mean Square Error (RMSE), and Mean Absolute Percentage Error(MAPE). In a days and water table depth prediction phase Support, Vector Regressor, Deep Neural Network, and TabNet Regressor are used, and compare the results with our proposed Ensemble Number of Days (E-NOD) and Ensemble Water Table depth (E-WTD) Regressor model. E-NOD and E-WTD models achieved less MAE, RMSE, and MSE than other machine learning methods.

INDEX TERMS Applied machine learning, ensemble learning, voting classifier, deep tabular model.

I. INTRODUCTION

Cleanwater plays an essential role in achieving industrial and economic development, whether used for food production,

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

drinking, or domestic use. On earth, 97% of the water is salt water, and just 3% is fresh water. But two-thirds of the freshwater is frozen in glaciers and polar ice caps [1]. About 30% of the freshwater is below the earth's surface, known as groundwater [2]. Groundwater is the primary water source necessary for agriculture, irrigation, industrial activities, and drinking around the globe [3]. Groundwater is immensely important for maintaining the biodiversity of the region [4]. In the past, people used hand shovel to dig wells for groundwater acquisition, but this process was time taking and required more human efforts. Drilling machines are used to dig a bore well by drilling a borehole in the ground in search of water. The Percussion Drilling method is done by putting a 50 kilo gram of heavy cutting tool in the hole and it has low cost operation cost. The boreholes are drilled considering the water table at that particular drilling site. The water table is an underground boundary between the area where groundwater saturates in the soil surface. The availability of groundwater is influenced by the water table depth, which varies significantly across various regions. In some regions, the drilling process takes more time than other regions. Down the earth's surface are various layers and soil types. The subsurface soil layers and soil types exhibit specific chemical and physical properties. The soil composition closely relates to climates and geological and horological characteristics of that region area. For instance, some areas have soft underground soil layers, while others have a hard underground soil layer. The hardness level of soil layers ascertains the time and cost of resources required to drill a borehole for the extraction of scarce water resources. Consequently, borehole placement on a drilling site with hard soil composition renders expensive machinery, skilled workforce, and time budget compared to a soft underground soil layer.

Due to population growth, demands for freshwater water have increased [5]. To meet the global water demand a huge number of bore wells are being drilled, resulting in over-exploitation of scarce groundwater resources. Due to a surge in drilling operations a wealth of borehole drilling data is being generated. The high dimensional and dynamic borehole drilling data requires in-depth analysis and modeling [6]. Furthermore, the drilling industry is a multi-billion industry embodying highly skilled task forces, and heavy machinery involving massive budgets. Thus over and under-utilization of resources can be a cause of major loss to drilling companies. In the past few decades, advanced technologies have been employed to speed up the drilling process and to minimize the drilling time to reach the water table depth. Factors like soil hardness, water table depth, and number of days spent on the drilling process in certain regions are essential to be considered before starting the drilling process.

The predictive analytics help the drilling companies and hydro-geological resource managers in effective planning to estimate drilling cost and drilling resources in advance. Therefore there is a dire need to analyze the vast amount of data generated by the drilling process [7] to extract the information from the data. Due to technological advancement Machine Learning (ML) and Artificial Intelligence (AI) approaches are being widely adopted across many domains [8]. The machine learning algorithms are capable enough to learn from data, identify underlying patterns, and help make organization owners make informed decisions. Furthermore Machine Learning methods are highly effective at solving complex problems by mapping the spatial and temporal correlations by learning the patterns in time series data. For instance trend forecasting in the stock exchange, Computer-aided diagnosis systems, and text classification. To perform the analysis and extracting the information from the rich hydro geological data-sets, machine learning method are highly preferred due to their exceptional performance. In a article [9] employed machine learning techniques method for prediction of subsurface structures using well logging data. For sustainable management of water resources ground potential maps plays important role. Reference [10] use Naïve Bayes (NB) based ensemble model (integrate Naive Bayes with Bagging, Adaboost and Rotation Forest) for ground water potential classification on the dataset of Kon Tum Province, Vietnam. Reference [11] employed entropy, Gini and Ratio-based classification tree for ground water potential classification [add detail] in a mountainous region of Iran. In a study [12] employed several machine learning based approaches such as Artificial Neural Network (ANN), [13] employed Support Vector Machines (SVM) and [14] employed the Adaptive Neuro-fuzzy Inference systems (ANFIS) for prediction of ground water level. Deep learning frameworks are becoming popular recently because of their ability to handle large size data while producing better performing models. Several deep learning models such as Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) are well known to be highly efficient at managing the long term dependencies. For instance [15] proposed Long Short-Term Memory model for prediction of water-table depth using real dataset of Hetao Irrigation District China.

The depth of the water table, soil layer and number of days required for borehole drilling are varies from region to region. The resources for drilling are varied due to the hardness level of the soil. To meet the water needs, the development authorities launch several drilling projects for water extraction in particular regions. The allocated resources, budget, and time need to be taken into account beforehand and the concerned department. For the sake of that, expensive tests and site surveys are performed by drilling and hydro-geological experts in a specific area based on their experience. Unfortunately, there is no automatic way of doing so. Machine learning-based solutions can be employed to accurately predict sub-surface properties and hydro-geological characteristics such as soil type, water table depth, and time required to reach the water table in a specific region.

To address the aforementioned issues, we developed models based on SoTA machine learning techniques to provide

useful insights through predicting water table depth, soil layer and the number of days to facilitate groundwater resource managers and drilling companies. This proposed work develops an adaptive Ensemble Weighted Voting Classifier (EWV-SLC) to provide prediction results that assist management to take informed decisions. The voting strategy combines the knowledge of all the candidate classifiers and assign class labels based on their weighted contribution. The motivation behind proposing an EWV-SLC is that the ensemble learning classifiers is an better prediction model when we have a imbalanced class label. For water table prediction Random Forest (RF), Extreme Gradient Boosting (XGB) and Bagging Regressor as selected as candidate learners. While to predict the number of days, Decision Tree (DT), RF, and XGB as candidate learners. The experiments are performed on the real dataset comprising of borehole logs. To evaluate the proposed model we performed comparative analysis with baseline models including Support Vector Machine, TabNet, Deep Neural Network (DNN) and Deep Tabular models. land layer is an attribute that represents the layer of land at different depths. In the borehole log dataset, we have 7 land layers including Gyeongam Formation, Landfill Layer, Ordinary rock formation, Sedimentary layer, soft rock layer, Weathered rock layer, Weathered soil layers. The land layer class label is suffering from the imbalanced class problem. To solve the imbalanced class problem, Synthetic Minority Oversampling Technique (SMOTE) data re-sampling technique is used.

The key use cases of the proposed study are listed below:

- Environmental Planning: The proposed model can provide assistance to an environmental planning organization in selecting suitable locations for urban planning and infrastructure development
- Assessment of Environmental impact: Environmental impact assessments are crucial before starting the development of an industrial project. The proposed study assists decision-makers in minimizing the consequences by assessing the potential of the soil layer.
- Water Resource Management: The proposed study can assist the water resource manager to manage the water resources for water conservation strategies, and ensure the sustainable use water of water by utilizing the water table prediction.

The key contributions are listed below:

- Applied enhanced data pre-processing and feature engineering techniques on raw data to improve the quality of the data and make it more suitable for the machine learning models.
- Development and Integration of the bagging and boosting ensemble technique with k Nearest Neighbor (KNN) using weighted voting strategy to develop a land layer prediction model for optimal planning of groundwater extraction schemes.
- Development and integration of bagging and boosting ensemble technique with Decision Tree(DT) using ensemble voting strategy to predict number of days for optimal planning for water pumping schemes
- Development of water table depth prediction model for effective groundwater resource management by combining the bagging and boosting techniques through the use of an ensemble voting procedure.
- Comparative analysis with the best available deep learning models for drilling-process prediction to verify the effectiveness of the proposed model.

The rest of the paper is organized as follows. Section II discuss a detailed review of the existing work of various researcher for water pumping and resource management. The proposed methodology of our proposed work is presented in section III. Section IV discussed the results of our proposed work. The conclusion is presented in section V.

II. LITERATURE REVIEW

This section discusses the existing work done for effective management drilling and groundwater resources. The process of effective management of groundwater and drilling resources refers to systematically and cooperatively utilizing, protecting, extracting and developing groundwater resources for the long-term use. Therefore management and optimization of drilling and groundwater is essential to ensures a secure and sustainable water supply, Protects aquatic environments against depletion and contamination of groundwater resources. Clean and safe access to groundwater resource not only helps agriculture and business but is imperative to economic and social advantages over the long run. In addition to that improving the efficacy and efficiency of the drilling operations maximize water extraction while reducing drilling costs and environmental implications. In this regards several researchers have devised solutions for optimal resource planning of water pumping schemes. The decision-making process pertaining to groundwater and drilling optimization can be greatly aided by the application of machine learning. The authors in [16] developed a hybrid model using Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and RF for classification of drilling states in real-time. In [17] authors employed machine learning techniques for the risk assessment of ground water contamination occurrence using a Support Vector Machine (SVM), Multivariate Discriminant Analysis (MDA), and Boosted Regression Tree considering a dataset of 102 wells. In [18] the researcher provided an overview of the machine learning methods for drilling optimization and real time analysis of drilling parameters.

A. MACHINE LEARNING FOR WATER LEVEL PREDICTION

Machine Learning techniques are widely used for planning and managing hydrological resources, specifically the estimation of groundwater parameters. Water level estimation is essential for efficient groundwater management and achieving sustainable development goals. In [19] the authors estimated the water level using temperature and monthly mean precipitation using Multiple Linear Regression (MLR) and ANN. Similar work have been done in [20] to predict the groundwater level in the Reyhanli region of Turkey using

Artificial Neural Networks (ANN) and M5Tree models. The authors also modeled the impact of monthly average precipitation and temperature on the groundwater level. Reference [21] used rainfall and sea level data to forecast the groundwater level using RNN and LSTM model for flood management in coastal area. The reviewed works on Groundwater modeling suggest that LSTM achieved superior performance compared to RNN. Conventional machine learning-based solutions are also developed by several researchers such as Support Vector Machine, Random Forest, and K Nearest Neighbor are used in the literature. Support Vector Machine (SVM) is a machine learning technique that can be used in both classification and regression problems. SVM can be easily used with categorical and continuous multiple features. K Nearest Neighbor (KNN) is a supervised learning algorithm that can solve classification and regression problems. For instance, article [22] proposed a hybrid model based on K-Nearest Neighbor (KNN) and RF for water level prediction using solar radiation, daily mean temperature, precipitation and daily maximum solar radiation. To forecast the ground water level [23] utilized Extreme Learning Machine (ELM) and SVM model. Experiments are conducted on seasonal variables including temperature, rainfall, evaporation and transpiration.

B. MACHINE LEARNING FOR GROUND POTENTIAL CLASSIFICATION

Modeling groundwater potential is highly imperative for effective groundwater resource management. Article [24] proposed an ensemble learning model based on Logistic Regression, namely Random Subspace Logistic Regression (RSSLR), Dagging Logistic Regression(DLR), Cascade Generalization Logistic Regression (CGLR) and Bagging Logistic Regression (BLR) for potential groundwater mapping in the province of Vietnam. The proposed work utilized environmental factors as independent variables. The results of the study proved that DLR achieved highest superior performance compared to CGLR, RSSLR and BLR. Ensemble learning frameworks are known to improve the performance of the model by overcoming the limitations of weak learners. In [25] ensemble learning scheme is developed for modeling groundwater. The ensemble framework is built using J48 DT, Rotation Forest, Bagging, Dagging, Random subspace, and AdaBoost. The proposed work considered sixteen groundwater such as slope,topographic wetness index,elevation, distance from river network and elevation etc as independent variables for modeling. The results of the study suggest that RF-J48 achieved the highest AUC score of 0.797 among others. The authors in [26] proposed a RF-based Random Subspace classifier for mapping groundwater potential in Kurdistan province of Iran. Experimental results show that RF has a very high predictive power compared to Logistic Regression and Naive Bayes. In [27] the authors compared the performance of ensemble models on ground potential estimation. The proposed work developed Bagged CART,Random Forest, Boosted generalized additive model and AdaBoost, is applied on ground potential factors dataset such as groundwater productivity data and other groundwater potential conditioning factors. The prediction accuracy of RF was 86 % that is highest as compared to other methods.

C. MACHINE LEARNING BASED WATER TABLE PREDICTION

Groundwater table prediction plays an essential role in the planning and management of ground resources. Reference [28] compared the performance of RF and XGB in forecasting the water table depth for cranberry field farms in Canada. Experimental findings state that XGB achieved better results than RF for water table depth forecasting. In [29] authors developed ANFIS based model to forecast the groundwater table.

The fluctuation of water table depth occurs due to seasonal and environmental changes. To investigate the groundwater behavior [30] proposed Fuzzy Logic, Radial Basis Function Neural Network (RBFN), and Co-Active NFIS-based solution. Accurate modeling of groundwater resources is important for the management and planning of hydro-geological resources. In an article [31] proposed ANN model for seasonal ground water table depth prediction while Genetic Algorithm (GA) is used to optimize the weights of ANN. Reference [32] employed ANN and SVM for ground water forecasting.

Reference [33] devised a solution to monitor the long-term trend of groundwater table in the northwest region of Bangladesh using data from 350 wells. In a article, [34] developed Support Vector Regressor (SVR) and Controlled Auto Regressive Ridge Regression (CAT-RR) for groundwater table forecasting. Similar work is done by [35] to predict the groundwater table to evaluate the sustainability of groundwater resources using SVR and observe performance gains.

D. MACHINE LEARNING FOR DRILLING RATE OF PENETRATION PREDICTION

The drilling rate of penetration prediction is adapted to optimize drilling performance. In a study [36] used ANN, SVM, and Hybrid Multi-Layer Perceptron for drilling rate of penetration prediction. A hybrid ANN with a Simulated Annealing (SA), Invasive Weed Optimization Algorithm, Firefly Algorithm (FA), Shuffled Frog Leaping Algorithm, and Standard Back-propagation to learn the weights for drilling rate index estimation is proposed in [37]. The experimental results demonstrated that ANN with SA achieved noteworthy performance among all. Reference [38] used ANN for the Rate of penetration prediction. The authors optimized the weights of ANN using a self-adaptive differential equation. Similarly, in [39] the authors employed ANN to predict penetration rate. To fine-tune the ANN parameters, Artificial Bee Colony (ABC) is employed.

Reference [19] perform groundwater level prediction using temperature and monthly mean precipitation factors. Article

[24] classified the ground potential by making use of environmental factors. In [30] the authors predicted the water table using environmental and seasonal factors. In another study [31] ground water table depth is predicted using seasonal factors. The review of existing and current methods applied to optimizing drilling operations and managing groundwater resources suffers from the following limitations. Firstly Groundwater systems are notoriously difficult to anticipate because of their complexity and the fact that they are constantly changing in response to pumping, recharge, and other stratigraphic variables. The existing studies have not taken into account the subsurface lithologies and startigraphic uncertainties into account. Further the groundwater data is often collected by multiple organizations using diverse methodologies, making it challenging to compare and interpret data from different sources due to the lack of consistency in data collection and reporting. Predictive groundwater modeling relies heavily on the quality of the data used to make predictions. Inaccurate predictions, ill-informed decisions, and wasteful use of resources are all possible outcomes of low quality data. However, better data can lead to better groundwater modeling, better decisions, and better long-term sustainability and efficiency in managing groundwater supplies. The accuracy and reliability of models that predict hydro-logical and lithological attributes is often compromised due to a lack of proper validation with observed data. Moreover, in the previous studies, the researcher proposed a method for groundwater table prediction. But they all predicted groundwater table using environmental and seasonal variables.

The quest of devising a machine learning-based solution to model the down-hole environment automatically is ever followed. A robust solution is required that can detect soil layers and suggest the time and resources required to drill and extract groundwater to reach a specific water table depth using geological features. To this aim, we developed a robust model to predict the water table depth, soil layer, and drilling time to reach a specific water table depth. For improved model performance and accurate results, we used a real borehole dataset provided and collected by a research organization. The proposed model harnesses the lithology and drilling features to train and test the proposed models; geological layer name, latitude, longitude, altitude, starting depth, ending depth, and soil color. Our developed model is scalable to dynamically varying down-hole environments. The developed system allow drilling companies and water boards to maximize efficiency by forecasting water needs based on historical data. The proposed predictive models can be used to determine the optimum water sources for drilling operations, taking into account formation structure parameters (soil layer properties, water level, drilling depth) to ascertain the groundwater availability and cost of drilling operations. The proposed system helps the management teams to take informed decision based on Prediction results. The predictive models can offer valuable insights that inform decision-making, enabling water boards and drilling firms to make data-driven water usage and management decisions.

III. PROPOSED METHODOLOGY FOR WATER RESOURCE MANAGEMENT

This section discusses the methodology of our proposed work for groundwater and drilling resource optimization. management and Algorithm 1 shows the steps of the proposed method. The first section, discussed the detail of the data set. The pre-processing steps are discussed in section two. The proposed models for water resource management are presented in section three. Finally, we evaluate the performance of our proposed methods. The detail of evaluation metrics is discussed in section four. Figure 1 describes the phases of our proposed work for water resource

A. DATA-SET

The experiment is performed on the borehole log data set. The borehole log data set contains several attributes related to drilling points. This data-set set contains 9287 instances of 1987 unique borehole logs. The time required to dig a borehole is known as drilling time. The instances represent unique occurrences of drilling time, or the time spent drilling at a particular location for extraction of groundwater. In other words, each instance of drilling time represents the amount of time spent on a particular borehole. Groundwater extraction entails digging boreholes to reach subsurface water sources. Depending on criteria such as Drilling depth, water table, and geological conditions(soil layer and composition), the drilling machine may spend varying lengths of time on each borehole. By keeping track of the time required to drill each borehole, drilling companies can collect data to enhance their drilling procedures and increase productivity. Table 1 describes the detail of the borehole log data set's feature detail.

The borehole log data include the features related to drilling points such as geographic coordinates, soil color, soil layer, borehole log ID, ending and starting depth, the thickness of the layer, and groundwater level. A land layer is a rock unit under the ground surface. The land layer can be classified into different rock layers. The land layer is a target attribute that contains seven unique classes of rock layers, including, the Gyeongam Formation, Landfill Layer, Ordinary rock formation, Sedimentary layer, soft rock layer, Weathered rock layer, and Weathered soil layer. In the raw borehole log dataset, there are multiple records related to each drilling point. The records pertaining to a single drilling point contain related information to each point. There are multiple soil colors and land layers encountered during the drilling process for groundwater extraction. During drilling multiple soil colors and land layers are encountered, and this information can be used to better comprehend the underlying conditions. The various soil colors and land layers can provide information about the existence of groundwater by

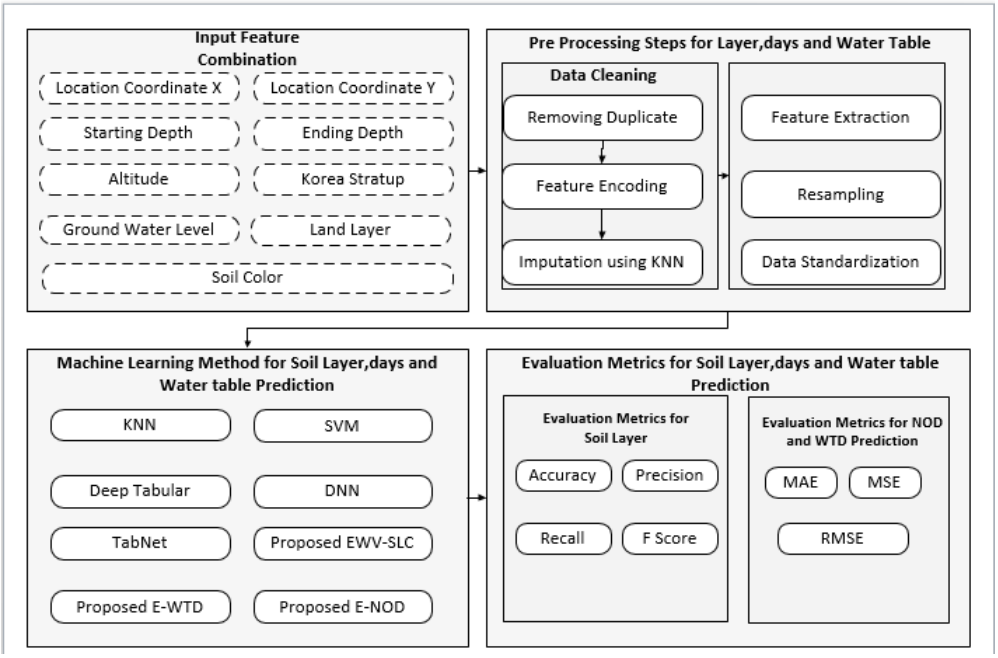

FIGURE 1. Proposed methodology diagram for prediction of soil layer, water table, and days prediction.

Algorithm 1 An Adaptive Ensemble Learning Assisted Cost-Effective Percussion Drilling for Water Boreholes

Data: Drilling Dataset $data = (x_1, x_2, x_3, \ldots x_n)$
Result: Prediction of Soil Layer, Days and Water Table Depth
initialization;
data ← (READCSV);
$Feature \leftarrow SplitFeature(data)$
for *each Feature* **do**
  data ← $removeduplicate(data)$;
  if *AlphaNumericFeatures* **then**
    data ← $FeatureEncoding(data)$
  $data \leftarrow KNNImputer(data)$
for *i from 1 to Data* **do**
  $data \leftarrow SMOTE(data)$
  $data \leftarrow Under\ Sampling(data)$
$Models \leftarrow InitilizeParameters()$
for *each* ***Models*** **do**
  $Model \leftarrow ModelTraining(data)$;
  $\hat{Y} \leftarrow ModelPrediction(DataWithoutLabel)$;
  Evaluate;
  $Accuracy \leftarrow TP + TN/TP + FN + FP + TN$
  $Precision \leftarrow TP/(TP + FP)$
  $Recall \leftarrow TP/(TP + FN)$
  $FScore \leftarrow 2 * (PR)/(P + R)$
  $MSE \leftarrow 1/m \sum_{i=1}^{m} (yi - \hat{yi})^2$
  $MAE \leftarrow \sum_{i=1}^{m} |(yi - \hat{yi})|$
  $MSE \leftarrow Sqrt(1/m \sum_{i=1}^{m} (yi - \hat{yi})^2)$;

indicating changes in soil composition. In the same way, through data analysis, we discovered that the number of days spent on each point and water table depth varies based on numerous factors, such as the depth of the borehole, the kind of soil and rock being drilled. On average, it may take a few days to drill a standard borehole, although more intricate

TABLE 1. Detail description of drilling data set's feature.

Feature Name	Description
X	X location coordinate represents the Longitude
Y	Y location coordinate represents the latitude
Altitude	Regions on the Earth's surface
Groundwater level	Represent the groundwater level
Starting depth	It represents the start depth of that specific day
Ending depth	It represents the end depth of that specific day
Starting thickness	It represents the starting thickness of the borehole
Land Layer	Represent the geological layers name such as Sedimentary layer, landfill layer, soft rock layer, weathered rock layer, and weathered soil layer.
Soil color	Represent the soil color like dark brown

or deeper boreholes may take longer based on strati-graphic uncertainty.

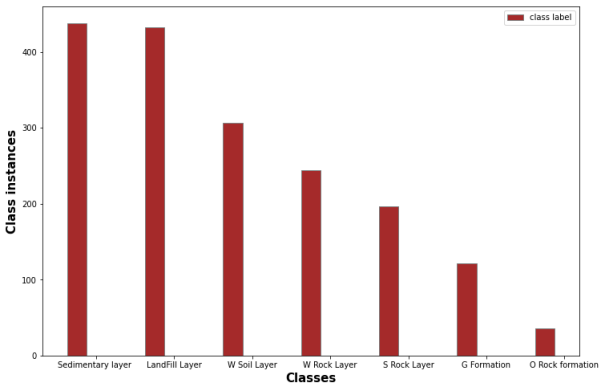

FIGURE 2. Frequency distribution of land layer class label: Seven land layers.

Figure 2 shows the frequency distribution of the class label. It shows that the ordinary rock formation layer is suffering from a class imbalance problem. during data pre-processing step, we discussed the solution of this problem later in the manuscript.

B. PRE-PROCESSING OF BOREHOLE LOG DATA SET

Data pre-processing is a process of cleaning and converting the raw data into a reliable format. The fundamental objective of data pre-processing is to prepare data for analysis and model training by cleansing, converting, and structuring it into a format appropriate for statistical analysis and machine learning algorithms. The objective is to increase the data's quality and utility by eliminating errors, inconsistencies, and outliers and guaranteeing that it can be easily examined and comprehended. Data pre-processing is a crucial phase in the data analysis procedure, as it has a substantial impact on the correctness and dependability of the results [40]. Therefore, it is necessary to pre-process the raw data to handle missing values and outliers. In this study, we perform several steps to pre-process the raw data.

1) REMOVING REPEATED VALUES AND CATEGORICAL FEATURE ENCODING

The Borehole log data set contains repeated borehole data points. We remove the repeated borehole log data points for consistency of data. Remove redundant data points by applying the Structure Query Language (SQL) 's group-by method. To perform the mathematical operation, it is required that data must be in numerical format. The feature encoder is used to convert the categorical data values into numerical ones. The employed method assigns a unique numerical value to each category of a feature. We perform label encoding by using the Scikit feature encoder.

2) IMPUTING MISSING VALUES

The objective of handling missing values is to estimate or replace missing data in a dataset so that it may be analyzed. Handling missing values is crucial because missing data might affect the precision and validity of an analysis's conclusions. There are numerous approaches for dealing with missing values, including as replacing them with the mean or median of the available data, utilizing predictive modeling to estimate missing values based on trends in the data, or simply deleting occurrences with missing values from the study. Depending on the nature of the data and the objectives of the study, the proper strategy for addressing missing values will vary [41]. We Pre-processed the borehole log data set to find out and remove the missing values. We find that the soil color attribute has some missing values. To fill in the missing values, we used the KNN imputation method to impute the missing value. The KNN imputation method uses distance measure to identify the neighboring point. The complete neighboring observations are used to estimate the missing values [42]. The missing value is imputed with the estimated value by the scikit imputer method considering the two neighbor points. The KNN imputation is performed using Eq.1.

$$dist(x, y) = Sqrt(weight * (distfrompresentcoord)^2) \quad (1)$$

where Weight is equal to the total number of coordinates divided by a number of present coordinates.

3) FEATURE EXTRACTION AND SELECTION FOR BOREHOLE LOG DATASET

Feature extraction is a process of combining variables into features, to reduce the number of features. In Feature engineering, the extraction of relevant features helps us to increase the performance of the predictive model [43]. In this study, we extract some features from the existing features to improve the performance of the prediction model. In the feature extraction step, first, we extract the total depth feature related to each drilling point. The total depth feature is

extracted using Eq.2.

$$TD = \sum_{i=1}^{n} ED - SD \tag{2}$$

where TD is the total depth, SD is the starting depth, ED is the ending depth and n is the number of instances of a borehole.

Multiple days are spent on each drilling point, to count the number of days spent on each drilling point pandas value count function is used. Eq.3 Describe the equation of a number of days calculation.

$$days = \sum_{j=1}^{m} Count(j) \tag{3}$$

The water table is an underground boundary between the area where groundwater saturates and the soil surface. The borehole log dataset has an ending depth and amount of water level. In this work, we will predict the Water table. We calculate the water table using Eq.4

$$WaterTable = TotalDepth - Amountofwaterlevel \tag{4}$$

When a drilling process starts and the point where groundwater is found is called water level and the complete depth of the borehole is called the total depth. The groundwater level is the height of the water table within an aquifer, which is the water-bearing layer of porous rock, sand, or gravel. For example, the drilling process stop drilling on 30 meters, and achieved water on 25 meters, so amount of water level is 5 meter and total depth is 30 meters. The total depth of each borehole is varied on different locations. Where Total depth is the total drilling depth of borehole and amount of water level is water achieved in each drilling point.

This study employs location coordinates X, Y, altitude, soil color, and depth features for the classification of a land layer. The water table prediction is performed using soil color, land layer, X, Y, and altitude. Total depth, X, Y, altitude, soil color, and land layer are used for a day's prediction.

4) SMOTE FOR RE SAMPLING

When the distribution of data points is biased or skewed toward some classes, it's called an class imbalance problem [44]. In borehole log data set, the distribution of data is skewed towards the sedimentary layer. Where ordinary rock formation layer is minority class. To solve the problem of imbalance class, this study re-samples the data points to balance the data. The process of adding or removing the data instance is called re-sampling. When a new data point is added into minority classes data point in the data, it's called am oversampling. Therefore, when a samples is removed from majority class to balance the data set, it is called under sampling [45].

Synthetic Minority Oversampling Technique (SMOTE) is used to re-sample the data points. In re-sampling process SMOTE select the data points that are close in the feature space, draw a line between the data points in the feature space and draw a new data point sample at a point along that line [46]. This study first re-sample the minority ordinary rock formation layer equal to the data points of sedimentary layer (majority class). Secondly, perform the under sampling in the majority class like sedimentary layer and landfill layer and over sampling in the other minority classes such as ordinary rock formation layer and gyeongam formation layer. Finally, over sample all classes equal to the sedimentary layer.

5) DATA STANDARDIZATION

Data Standardization is a process of converting the data into a uniform format. It transforms the data on the same scale. The standard scaler method is used for the standardization of data. It reduces the effects of the too large or too small values of features [47]. The formulation of standard scalar is described in Eq.5.

$$Z = X - \mu/\sigma \tag{5}$$

where Z is a new Scale value, X is a value that needs to be normalized, μ is a mean of distribution and σ is a variance of the distribution.

C. PREDICTIVE MODELING OF GROUNDWATER RESOURCE

1) CLASSIFICATION TECHNIQUES FOR GROUNDWATER RESOURCE

This study used Support Vector Machine [48] for the classification of soil layers. It separates different multiple classes (Soil Layers) by constructing a hyperplane into multi-dimensional space. It can minimize the error by iteratively generating an optimal hyperplane. Drilling Dataset is given as input to SVM model, and the kernelling trick is applied to transform the low-dimensional space into a higher-dimensional space. The reason to select SVM as a classifier is due to the nonlinear nature of drilling dataset. Because of the kernel method, SVM works better when applied to a nonlinear problems. This study uses RBF kernel and set the value of gamma parameter as 0.9. RBF kernel performed the transformation of feature space using Eq.6.

$$K(x, x') = exp(-\|x - x'\|^2/2\sigma^2) \tag{6}$$

where σ is a variance and $\|x - x'\|^2$is a Euclidean distance between two points.

This study gives input the Drilling raw dataset to a TabNet model [49] without any preprocessing and this model is trained using gradient descent. Figure 3 shows the working flow of the Tabnet model. This method chooses the features at each step using sequential attention. It performs instance-wise feature selection; which means that each row on the training dataset can be different. TabNet performs soft feature selection because feature selection employs single deep learning architecture. TabNet provides both local and global interpretability. TabNet model is trained on certain parameters for water table depth, soil layers, and days prediction.

The decision step is a hyperparameter that is used in model training. The large step size increases the learning capacity of

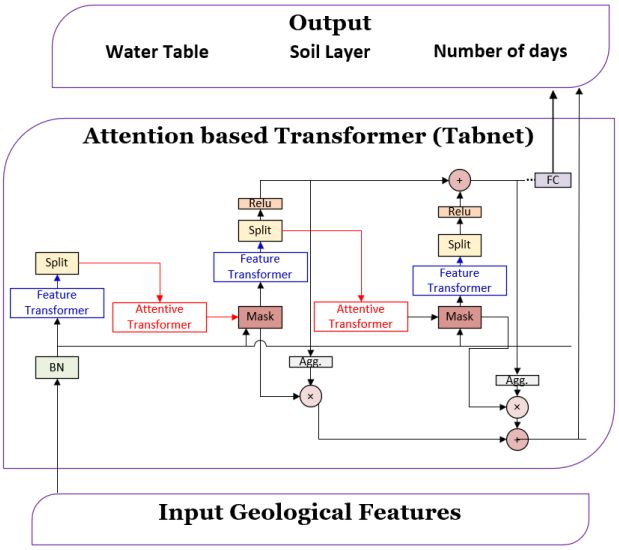

FIGURE 3. Architecture of Tabnet model for soil layer, water table, and days prediction: Input the geological features and it predicts the water table, soil layer and days.

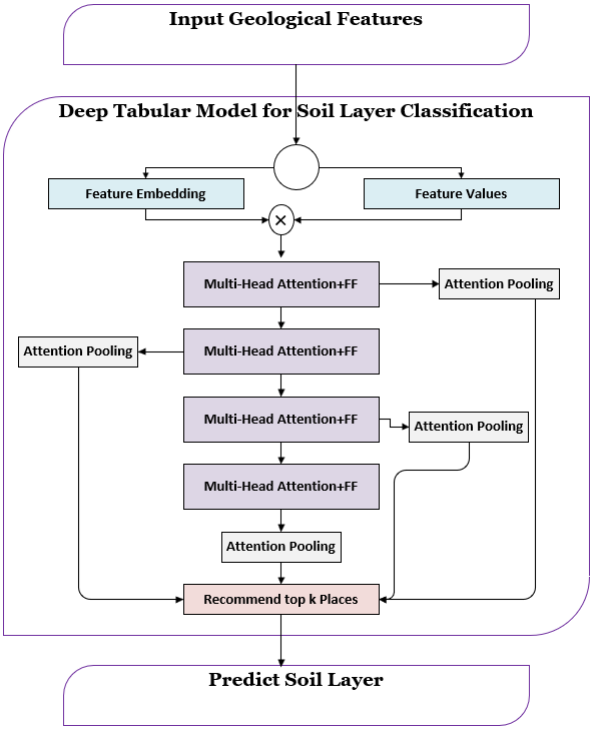

FIGURE 4. Architecture of deep tabular model for soil layer classification. Input the geological features and it predicts the soil layer.

the model; however on the other hand it increases training time. This study trained Tabnet with the value of step 3. The output of the decision step is combined by getting each step vote in the final classification, and votes are equally weighted. Attentive transformer includes the prior scales to know about each feature and how the previous step uses them. This attentive transformer derives the mask by using the previous features. The mask derives the explainability and is used to ensure that the model must focus on more relevant features. The sparsemax mask is used for each decision step. Generally, TabNet uses instance-wise feature selection, and features are selected for each input, meaning each prediction is performed on different features. This study used the Tabnet decoder module to fill in the missing values. Because we pass the raw data as input to TabNet.

The deep Tabular model's architecture is derived from TabNet [49], and AutoInt [50]. This method performs drilling feature set embedding, and multi-head self-attention like the AutoInt model, and the Pretraining part is derived from TabNet. Figure 4 shows the working flow of the deep tabular model. In a deep tabular model, the name of the drilling feature set is fed to the embedding layer and then multiplied with drilling feature values. The model used a point-wise feed-forward layer and a sequence of multi-head attention blocks. These attention blocks model the interaction between the drilling feature sets while the attention pooled skip the connection to get a single vector from the drilling feature embedding set.

K Nearest Neighbor [51] belongs to the class of lazy learners because KNN does not learn from the drilling feature training set immediately, alternatively, it stores the available dataset. KNN performs the action on the drilling dataset at the time of classification. KNN is a non-parametric method; on underlying data it does not make any new assumptions to classify soil layer samples. It calculates the similarity between available cases and new cases. It assigns the category of new cases that is most like the category of available cases based on the similarity. There is three most common type of distance metrics that can be used to calculate the distance between data points, such as Minkowski, Euclidean, and Manhattan distance. K is the main parameter of KNN that is used to decide the neighbors for calculating the distance. In our work, we used Minkowski distance metrics as distance metrics and set k equal to 2.

Soil layer class label has an imbalanced class label, and for imbalanced class distribution ensemble voting classifier [52] is an optimal prediction model. The single prediction model can be biased towards a majority class and that reduces the generalization of the model. In contrast, the Ensemble voting classifier has multiple base learners and in training it merges the knowledge from each base learner that increased the generalization of the model in a prediction phase. In a regression problem such as the number of days and water table prediction, an outlier can exist that effect the prediction result and a single classifier can be affected by an outlier. However, the Ensemble average regressor is robust of an outlier because it merges the knowledge from multiple base learners. First, we applied the different machine learning classifiers such as Random Forest, Gradient Boosting Decision

tree, K Nearest Neighbor, and others on drilling data-set for soil layer prediction. Afterward, we selected the candidate learner for ensemble learning who has high prediction accuracy and low error for the water table, days, and soil layer. To achieve better prediction results we performed parameter tuning of each model. Thus Grid search is used for hyper-parameter tuning of each candidate learner. The accuracy of each model is considered as a weight for each classifier. The higher prediction accuracy translates into to higher weight assignment.

This proposed work study selected the following model; Random Forest (RF) [53], Gradient Boosting (GB) [54], XG Boost (XGB) [55], and K Nearest Neighbor (KNN) classifiers for a weighted voting wrapper. In this proposed model merge the knowledge of RF, GB, XGB, and KNN for layer classification. Figure 5 shows a block diagram of Ensemble Weighted Voting Ensemble Soil Layer Classifier (EWV-SLC) for soil layer classification. The purpose of selecting the RF as a base learner is that it can automatically balance the dataset in case of an imbalanced class. Where Gradient Boosting is also suitable for performing the classification in the case of an imbalance classification problem. The aim of selecting XGB is that it works well with structure data and achieves good prediction performance. KNN is an instance base method that works based on the neighbor's value. The patterns of soil color and layer are related to each other. That's why we select KNN as the candidate for weighted voting.

2) MACHINE LEARNING MODEL FOR DAYS AND WATER TABLE DEPTH PREDICTION

The water table and the number of days prediction is a regression problem. To predict the number of days and water table this study used Support Vector Regressor, TabNet, Deep Neural Network and develop an ensemble averaging model for days and water table depth prediction. Support Vector Regressor is trained using RBF kernel. For experimentation of DNN model we set the batch size value as 64, maximum epoch as 1000, Adam optimizer is used to optimize the weights, and linear activation function is used at the output layer to produce the output. This study develops a DNN model for the water table and the number of days prediction. In Deep Neural Network [56], we multiply the weights with drilling features input value and pass them to the activation function. The linear activation function produces the output. The equation of net input is described in Eq.7. In our work, we used the ReLU activation function in hidden layers and linear activation function at the output layer. The Water table and days prediction is performed using the linear activation function. The Adam optimizer is used to optimize the weights of the network. Adam optimizer used Eq.8 to update the weights.

$$a = \phi(wx + b) \tag{7}$$

W is a weight, b is a bias and x (drilling features) are an input vectors. ϕ is an activation function.

$$Wnew = Wold - \alpha Vdw^{c}orr \tag{8}$$

We have proposed the Ensemble Number of Days (E-NOD) regressor model. In our proposed model for a number of days prediction, we have merged the knowledge of the Decision Tree Regressor, Xtra Tree Gradient Boosting Regressor, and Random Forest Regressor. Figure 5 shows the block diagram of the proposed model for a number of days prediction. To predict the water table depth, we have proposed the Ensemble Water Table Depth (E-WTD) model based on, Random Forest Regressor, Xtra Tree Gradient Boosting Regressor, and Bagging Regressor. Figure 5 presents the block diagram of the E-WTD model for more clarity. The purpose of using the XGB regressor model is attributed to its better performance on tabular data.

Figure 5 shows the proposed EWV-SLC for soil layer classification, the E-NOD model for days prediction, and the E-WTD model for a Water Table prediction. Bagging Regressor is a meta-estimator that fits base regressors on each subset of the actual dataset, and it finally performs the individual predictions by averaging. A decision tree [57] is a rule-based model that split the dataset into smaller subsets and incrementally developed the tree. It is a famous entropy-based regressor that commonly performs better in the case of tabular data. Random Forest is a Bagging base method. In both proposed models (E-NOD) and (E-WTD) we combine the prediction from multiple base models. First, we trained base models on data and get the prediction results from all base models. Finally, E-NOD) and E-WTD makes a prediction that is the average of base-estimator

Give an input geological features set to all the candidate classifiers (KNN, RF, GB, and XGB) and train each candidate classifier on the features set for EWV-SLC. Each classifier performs prediction, and perform voting by assigning the weight to each classifier. By voting finally, we perform the final prediction (Soil Layer).

Give an input geological features set with days target attribute to DT, RF, and XGB Regressor for E-NOD. Each model is trained on the geological feature set and predicts the days that is required to reach the water table. Finally, we average the result of each regressor and predict in a specific region how many numbers of days are required to reach the water table.

Give an input geological features set with water table target attribute to Bagging Regressor, RF, and XGB Regressor for E-WTD. Each model is trained on the geological feature set and then predicts the water table. Finally, we average the result of each regressor and to produce the final water table prediction outcome.

D. EVALUATION METRICS

We evaluated the performance of prediction models using Accuracy, Precision, Recall, and F-score for soil layer classification. Mean Square Error, Root Mean Square Error, Mean Absolute Error, and Mean Absolute Percentage Error are used to evaluate the performance of the prediction model for water table days, prediction.

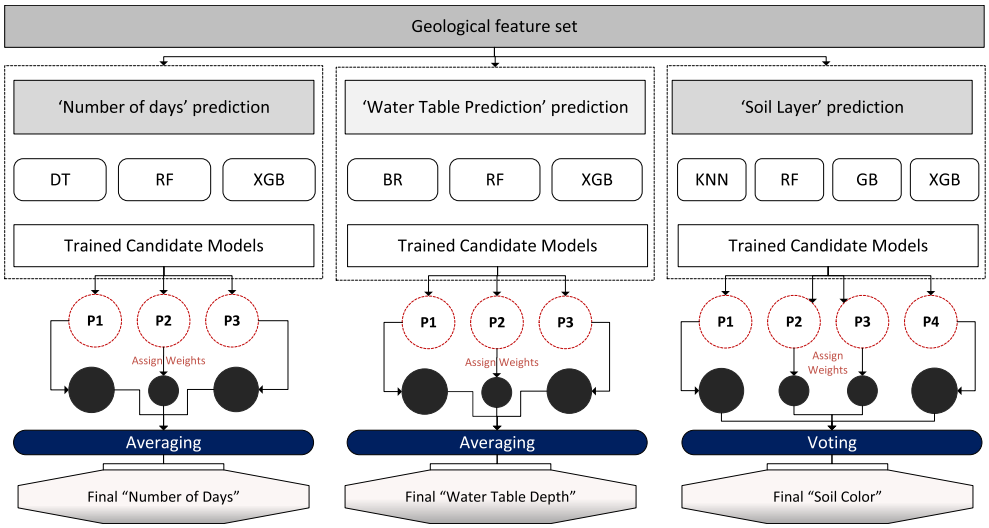

FIGURE 5. Proposed architecture of E-NOD, EWV-SLC and E-WTD model: Input the geological features and it predicts the water table, soil layer and day.

1) ACCURACY

Accuracy metrics describe how much our classifier predicts data points correctly from all data points. In Eq. 9, we write the formula of accuracy.

$$Accuracy = (TP + TN)/(P + N) \tag{9}$$

2) PRECISION

Precision is the probability that the classifier predicts a positive class correctly. In Eq. 10, we write the formula of precision.

$$Precision = TP/(TP + FP) \tag{10}$$

3) RECALL

The recall is the probability that the classifier predicts the actual class correctly. In Eq.11, we write the formula of recall metrics.

$$Recall = TP/(TP + FN) \tag{11}$$

4) F SCORE

F score is a performance measurement metric. It is a harmonic means of precision and recall. In Eq.12, we write the formula of F score metrics.

$$FScore = 2(PR)/(P + R) \tag{12}$$

In Eq. 12, P is precision, and R is a recall score.

5) MEAN SQUARE ERROR

Mean Square Error measures the difference between actual and predicted values. It tells us how much our predicted values are close to the actual value. In MSE we take the difference between the actual and the predicted value and square the difference. The lesser value of MSE determines closer fit and better performance of the model. MSE is calculated by using Eq.13.

$$MSE = 1/m \sum_{(i=1)}^{m} (y - ypred)^2 \tag{13}$$

where y is an actual value and ypred is a predicted value.

6) MEAN ABSOLUTE ERROR

Mean Absolute Error measures the difference between actual and predicted value by taking the absolute difference between actual and predicted value on the complete dataset. It is calculated by using Eq.14.

$$MAE = \sum_{(i=1)}^{m} |(y - ypred)| \tag{14}$$

7) ROOT MEAN SQUARE ERROR (RMSE)

Root Mean Square Error is a square root of MSE. RMSE determines the average distance that starts from the fitted line to the data points. RMSE is calculated using Eq.15.

$$RMSE = squareroot(\sum_{(i=1)}^{m} (y - ypred)^2)/m) \tag{15}$$

8) MEAN ABSOLUTE PERCENTAGE ERROR

Mean Absolute Percentage Error measures the percentage difference between predicted and actual values. It is calculated using eq.16.

$$MAPE = 1/n \sum_{(i=1)}^{m} (|y - ypred|)/y) \tag{16}$$

IV. RESULTS AND DISCUSSION

This section discusses the results of our proposed work for the water table, number of days, and soil layer prediction.

To predict the ground soil layer in the different regions at a specific depth this study applies a multi-class classification algorithm. This article performed the comparison between traditional Machine Learning methods such as SVM and KNN and Deep Learning-based models including Tabnet and Deep Tabular with the proposed Ensemble Weighted Voting Soil Layer Classifier (EWV-SLC) model. This study monitors the impact of KNN missing value imputation and SMOTE resampling technique for soil layer classification. Table 2 shows the classification result without imputation and resampling. To deal with the missing values this study imputes the missing values based on a distance-based method and ensures the better impact of this method in prediction performance. In Table 3, the results of the machine learning method with KNN imputation are presented. The soil land layer class label has imbalanced class distribution and to balance the class distribution of the soil layer this study used SMOTE re-sampling method because imbalanced distributions cause a problem of a classifier's biasedness toward the majority class. Table 4 shows the result of ML method with SMOTE minority class resampling and KNN imputation. In Table 5 describes the results of ML method with under and over-sampling and KNN imputation. Table 6 discussed the results of the machine learning method with missing value imputation and SMOTE oversampling resampling for soil layer classification. The prediction of the water table and the number of days is performed with machine learning models such as SVM and Deep Learning models such as Tabnet and Deep Neural Network. We compare the performance of these models with our proposed Ensemble model for the water table and number of days prediction. Table 7 and Table 8 show the result of the water table depth prediction without and with the imputation of missing values, and in Table 9 and Table 10 we discussed the results of a number of days prediction without and with the imputation of missing values.

A. SOIL LAYER CLASSIFICATION

In the borehole log data set, there are 7 land layers. To solve this multi-class problem, two deep learning models Tabnet and Deep Tabular are applied to predict the land layer in different locations. Empirical investigations suggest that the prediction performance of SVM and KNN is better as compared to the deep learning model. Moreover the proposed Ensemble Weighted Voting Soil Layer Classification (EWV-SLC) outperformed the baseline schemes SVM and KNN in terms of accuracy and F score.

TABLE 2. Experimental result of machine learning models without imputation and Resampling for soil layer classification.

Method	Accuracy	Precision	Recall	F Measure
KNN	77.3%	77.82%	77.3%	77.2%
SVM	69.99%	71.1%	69.9%	69.54%
Tab net	64.7%	64.7%	64.7%	64.7%
Deep Tabular	74%	69%	68%	68%
EWV-SLC	79.74%	80.23%	79.78%	79.56%

Table 2 shows the results of the machine learning classifier for soil layer classification without the missing value imputation and resampling. Table 2 shows that the KNN performed better in terms of accuracy as compared to Tabnet and the deep tabular model. While the prediction performance our proposed (EWV-SLC) model in terms of Accuracy, Precision, Recall, and F score are superior as compared to other models.

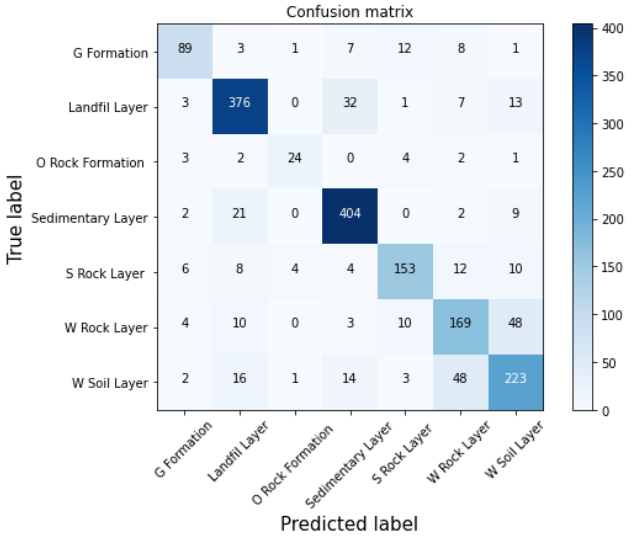

FIGURE 6. Confusion metrics of the proposed model for soil layer classification without imputation of missing value and Resampling.

Figure 6 shows the confusion metrics of our proposed model for the classification of soil layer without KNN imputation. It shows that the precision of the proposed model is 80.23% and recall is 79.78%.

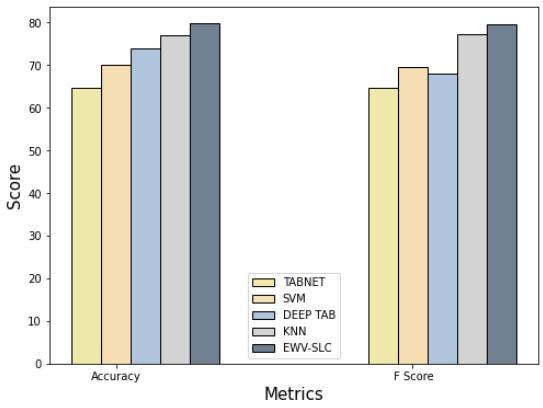

FIGURE 7. Comparison graph of machine learning method for soil layer classification based on a accuracy and f score metrics.

Figure 7 shows the comparison graph of the machine learning method for soil layer classification in a specific region based on accuracy and F score metrics without missing value imputation. It demonstrates that the accuracy and F score metrics of the EWV-SLC model are high as compared to other machine learning methods.

TABLE 3. Experimental results of machine learning using the KNN imputation and without Resampling for soil layer classification.

Method	Accuracy	Precision	Recall	F Measure
KNN	77.75%	78.35%	77.75%	77.6%
SVM	71.8%	72.63%	71.7%	71.44%
Tab net	64.9%	64.8%	64.92%	64.93%
Deep Tabular	74.5%	69.7%	68.34%	69.04%
EWV-SLC	80.5%	80.6%	80.45%	80.3%

Table 3 presents the results of the machine learning classifier for soil layer classification with the KNN missing value imputation and without the resampling. According to the results of Table 3, the accuracy of KNN is better as compared to TabNet and the deep tabular model. However, the comparative analysis of the proposed model (EWV-SLC) suggests that the model classifies the land soil layer with a high F score. Table 3 demonstrate that our missing value imputation technique worked well for imputing the soil color missing value. Due to the missing value imputation technique, the accuracy is increased almost to 0.67%.

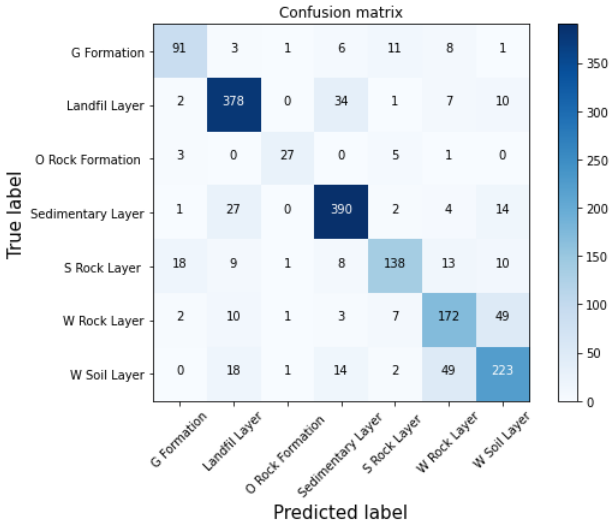

FIGURE 8. Confusion metrics of the proposed model for soil layer classification with KNN imputation and without resampling.

Figure 8 shows the confusion metrics of our proposed model for the classification of soil layers with KNN imputation and without resampling. It shows that the precision of the proposed model is 80.6% and recall is 80.45%.

Figure 9 depicts the graph comparing the machine learning approach for soil layer classification in a particular region with KNN imputation. The performance graph demonstrates that the accuracy and F score of the EWV-SLC model are superior to those of other machine learning techniques.

Table 4 shows the results of the machine learning classifier for soil layer classification with Smote minority class resampling and KNN missing value imputation. In SMOTE minority class resampling we oversample the minority class sample equal to the majority class. Table 4 demonstrates that our proposed model (EWV-SLC) classifies the land soil

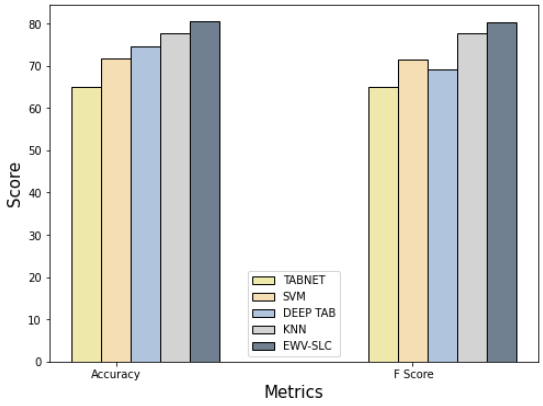

FIGURE 9. Comparison Graph of Machine learning method for soil layer classification with missing value imputation based on an accuracy and f score metrics.

TABLE 4. Experimental results of machine learning model using the KNN imputation and SMOTE minority class oversampling for soil layer classification.

Method	Accuracy	Precision	Recall	F Measure
KNN	82.5%	82.66%	82.5%	82.34%
SVM	76.07%	75.94%	76.1%	75.24%
Tab net	70.5%	70.5%	70.5%	70.5%
Deep Tabular	77.41%	77.52%	77.48%	77.42%
EWV-SLC	84.33%	84.49%	84.34%	84.2%

layer with high F score metrics that are highest as compared to other methods. Table 4 conclude that oversampling the minority class EWV-SLC model improved the almost 3% results in term of accuracy and F score.

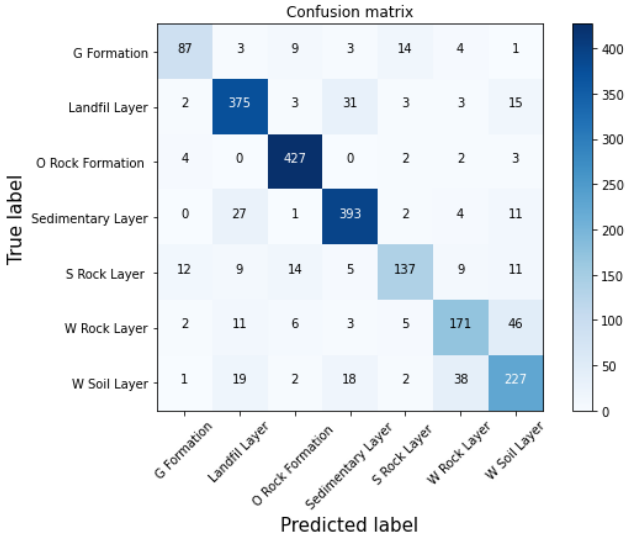

FIGURE 10. Confusion metrics of the proposed model for soil layer classification with KNN imputation and SMOTE Minority class resampling.

Figure 10 shows the confusion metrics of the EWV-SLC model for the classification of soil layer with KNN imputation and minority class oversampling. It shows that the

precision of the proposed model is 84.49% and recall is 84.34%.

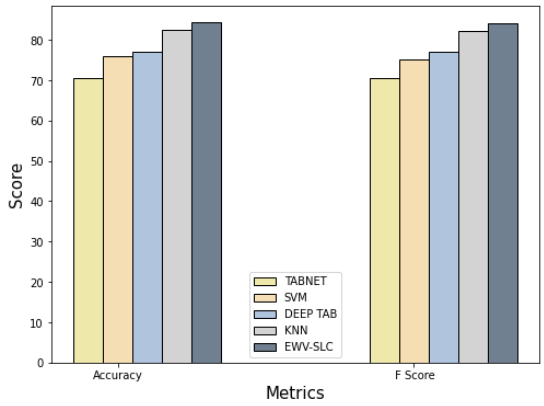

FIGURE 11. Comparison Graph of Machine learning method for soil layer classification with missing value imputation and Minority Class resampling based on accuracy and f score metrics.

In Figure 11 we present the performance of the machine learning method in terms of accuracy and F score. By applying the minority class resampling the results of EWV-SLC and other ML methods increased in terms of accuracy and F score.

TABLE 5. Experimental results of machine learning model using the using KNN imputation and with Smote under and oversampling for soil layer classification.

Method	Accuracy	Precision	Recall	F Measure
KNN	82.34%	82.6%	82.34%	82.28%
SVM	74.63%	75.66%	74.63%	74.43%
Tab net	68.7%	68%	68.7%	68.1%
Deep Tabular	76.3%	76.43%	76.38%	76.31%
EWV-SLC	84.72%	85.18%	84.72%	84.94%

In Table 5 we present the results of our proposed model, traditional machine learning, and Deep Learning Classifier for soil layer classification with SMOTE over and under-sampling. In oversampling, we add more samples to the data set, and we remove samples in under-sampling.

The Proposed model performed better in terms of accuracy and F score as compared to another method. By applying the SMOTE over-sampling and under-sampling the result of our proposed model is improved from 84.2% to 84.72%.

Figure 12 shows the confusion metrics of the EWV-SLC model for the classification of soil layer with KNN imputation and SMOTE over and under-sampling. The result verifies the efficacy of the proposed model in terms of precision and recall scores of 85.18% and 84.72% respectively.

Figure 13 shows the performance of the machine learning method in a bar chart with over and under-resampling for soil layer classification in a specific region. The performance of the EWV-SLC model is high as compared to another method in terms of accuracy and F score.

In Table 6 the results of the machine learning classifier are presented with SMOTE majority class oversampling.

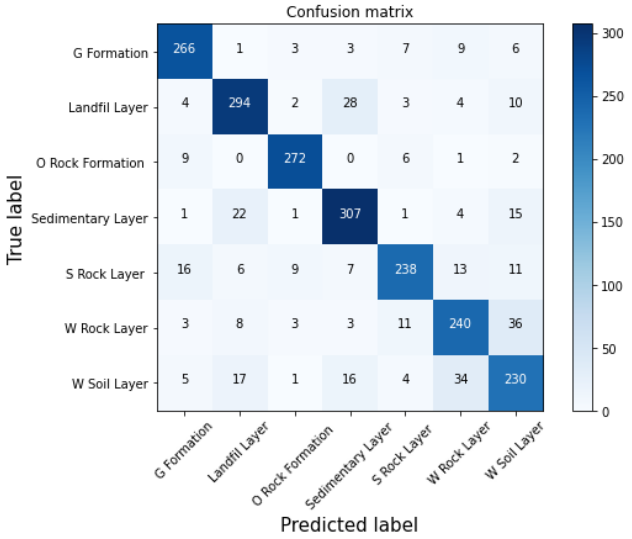

FIGURE 12. Confusion metrics of the proposed model for soil layer classification with KNN imputation and SMOTE under and oversampling.

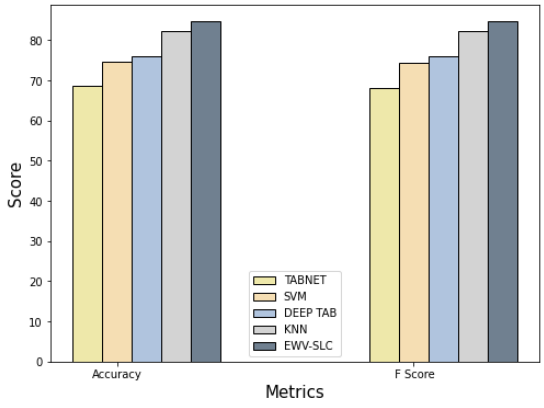

FIGURE 13. Comparison Graph of Machine learning method for soil layer classification with missing value imputation and over and under resampling based on a accuracy and f score metrics.

TABLE 6. Experimental results of machine learning model using the using KNN imputation and with Smote oversampling for soil layer classification.

Method	Accuracy	Precision	Recall	F Measure
KNN	83.38%	83.01%	82.34%	82.04%
SVM	78.57%	79.78%	78.57%	78.6%
Tab net	73.9%	74%%	73.9%	73.7%
Deep Tabular	77.5%	77.3%	77.81%	77.6%
EWV-SLC	89.11%	89.48%	89.11%	89.13%

In SMOTE majority class oversampling we over-sample the data points of all class labels equal to the majority class label. EWV-SLC achieved 89.13% accuracy that is the highest among all the methods. By applying the majority class oversampling, the F score of EWV-SLC is improved from 84.94% to 89.13%.

Figure 14 shows the confusion metrics of EWV-SLC model for the classification of soil layer with KNN imputation and

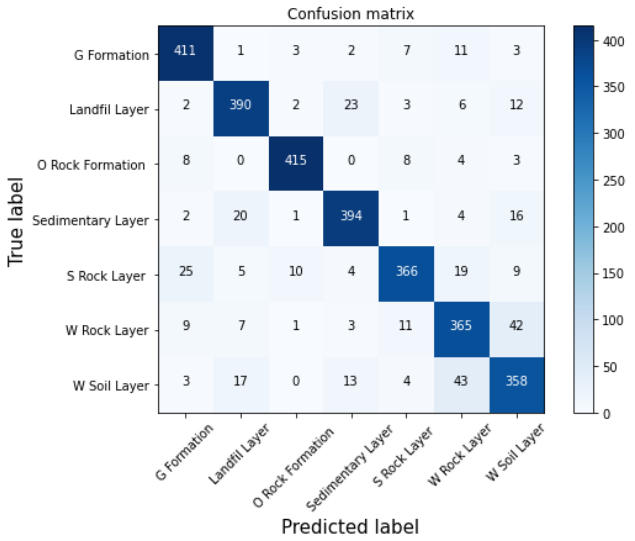

FIGURE 14. Confusion metrics of proposed model for soil layer classification with KNN imputation and SMOTE oversampling.

SMOTE Oversampling. It shows that the model precision of the proposed model is 89.48% and recall is 89.11%.

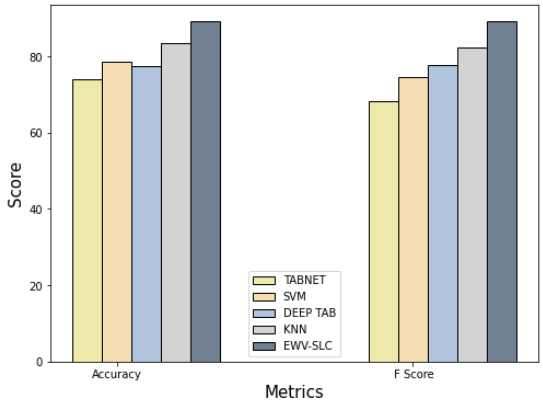

FIGURE 15. Comparison Graph of Machine learning method for soil layer classification with missing value imputation and SMOTE majority class oversampling based on a accuracy and f score metrics.

Figure 15 shows the comparison graph of the machine learning method for soil layer classification in a specific region with SMOTE majority class oversampling. It concludes that applying the majority oversampling EWV-SLC model achieved the highest result in terms of accuracy and F score.

B. WATER TABLE PREDICTION

In this subsection, we discussed the results of water table depth prediction. The water table is a regression problem, we predict the depth of the water table using Tabnet, DNN, SVR. To predict the accurate depth of water table we proposed our Ensemble Water Table depth prediction model (E-WTD). E-WTD model predicts the water table depth with less MAE error as compared to other models.

TABLE 7. Experimental result of machine learning models for Water Table Prediction without imputation of missing values.

Method	MAE	MSE	RMSE	MAPE
Tabnet	5.77	69.99	8.37	0.41
DNN	4.084	40.3	6.35	0.38
SVR	6.89	112.98	10.63	0.47
E-WTD	3.02	28.65	5.35	0.34

Table 7 describes the results of machine learning method for the prediction of water table depth without imputing the missing value. The Mean Absolute Error of the E-WTD model is less as compared to SVR, DNN, and TabNet. It demonstrates that the E-WTD model predicts the error WTD with less error.

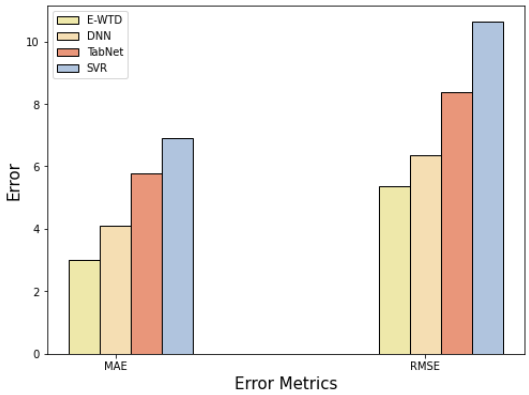

FIGURE 16. Comparative Analysis of machine learning model for Water Table Depth Prediction based on an MAE and RMSE metrics: without the KNN imputation.

In Figure 16 we present the MAE and RMSE of the machine learning method in bar chart for prediction of water table depth in a specific region. It shows that the error of the E-WDT model is less as compared to other method in predicting the depth of the water table.

TABLE 8. Experimental result of machine learning models for Water Table Prediction with the imputation of missing values.

Method	MAE	MSE	RMSE	MAPE
Tabnet	5.59	63.92	7.99	0.51
DNN	4.8	53.89	7.34	0.31
SVR	6.45	108.55	10.42	0.51
E-WTD	4.05	39.14	6.26	0.35

Table 8 describes the results of machine learning method for the prediction of water table depth with KNN imputer for the missing values. The Mean Absolute Error of the E-WTD model is less as compared to SVR, DNN, and TabNet. It demonstrates that the E-WTD model predicts the WTD with less error.

In Figure 17 we present the MAE and RMSE of the machine learning method in line chart for prediction of water

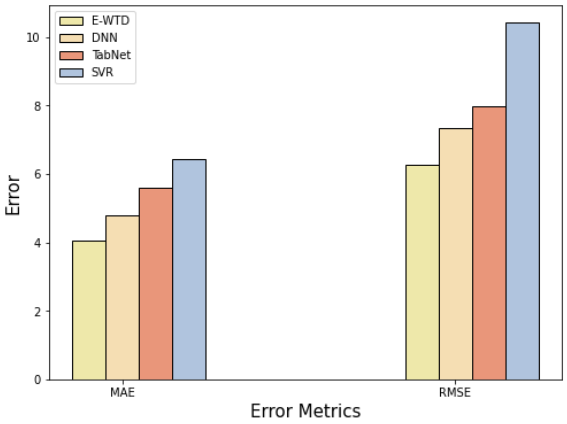

FIGURE 17. Comparative Analysis of machine learning model for Water Table Depth Prediction based on an MAE and RMSE metrics: with the KNN imputation.

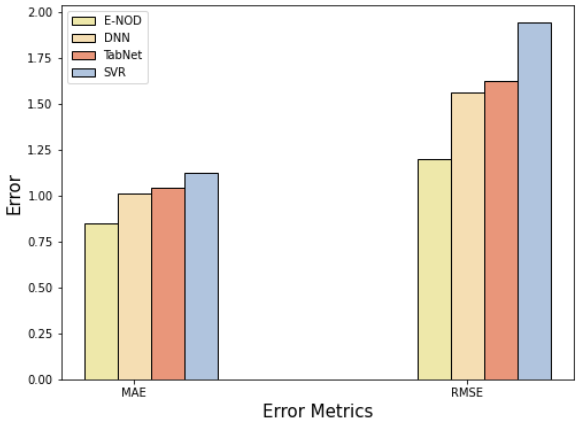

FIGURE 18. Comparative Analysis of machine learning model for days Prediction based on an MAE and RMSE metrics: without the KNN imputation.

table depth in a specific region without imputation of missing value imputation. It shows that the error of the E-WDT model is less as compared to other method in predicting the depth of the water table.

1) NUMBER OF DAYS PREDICTION

When we start drilling in a specific region if we already know the estimate of how many days is required to reach a water table in a specific region it helps us to estimate the cost and other resources. In this subsection, we discussed the results of days prediction. We apply Tabnet, DNN, SVR model for days prediction and compare the performance with our proposed model. Table 9 shows the results of the Number of Days prediction without imputation of missing values and Table 10 shows the results with imputation of missing values.

TABLE 9. Experimental result of machine learning models for Num of Days Prediction without Missing value imputation on a test dataset.

Method	MAE	MSE	RMSE	MAPE
Tabnet	1.04	2.64	1.62	0.236
DNN	1.01	2.45	1.56	0.23
SVR	1.126	3.76	1.94	0.25
E-NOD	0.849	1.46	1.20	0.21

In Table 9 the results of prediction models are discussed. It shows that our proposed E-NOD model predicts the number of days with less MAE error compared to DNN, Tabnet, and SVR. The takeaway of Table 9 is that the proposed E-NOD model is correctly predicting the required numbers of days are to reach the water table in a specific region.

Figure 18 describes the error of the machine learning model for the prediction of a number of days in a specific region. It demonstrates that the E-NOD prediction model predicts how many days are required to reach the water table depth with less MAE and RMSE.

In Table 10 the results of prediction models with KNN imputation are presented. It shows that our proposed E-NOD prediction model predicts the number of days with less MAE error as compared to DNN, TabNet, and SVR. Table 10 concludes that E-NOD correctly predicts how many days are required to reach the water table in a specific region.

TABLE 10. Experimental result of machine learning models for Num of Days Prediction with Missing value imputation using KNN imputer.

Method	MAE	MSE	RMSE	MAPE
Tabnet	1.31	4.49	2.12	0.27
DNN	1.18	3.79	1.55	0.25
SVR	1.87	11.26	3.35	0.26
Combined Voting	0.95	1.765	1.3	0.22

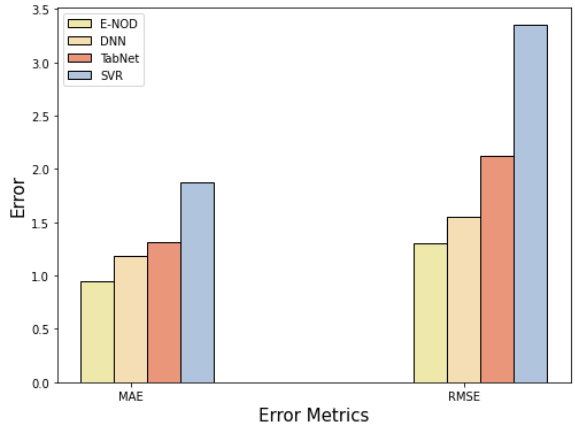

FIGURE 19. comparative Analysis of machine learning model for days Prediction based on an MAE and RMSE metrics: with the KNN imputation.

Figure 19 describes the error of the machine learning model for the prediction days in a specific region with missing value imputation. It demonstrates that the E-NOD model predicts how many days are required to reach the water table depth with less MAE and RMSE.

V. CONCLUSION

The proposed research study consists of three main modules. In the first module, the classification of a land layer is performed. In the classification module we first, balance

the class label by applying the SMOTE. In this module, we proposed an ensemble weighted voting classifier with KNN, bagging, and boosting techniques to predict the land layer for optimal planning for water pumping schemes. The accuracy, precision, recall, and F score of the EWV-SL Classifier are 89.11%,89.48%,89.11%, and 89.13% respectively which show increased efficiency. Our Precise prediction findings can assist in optimizing operations, hence decreasing resource wastage and boosting the productivity of drilling operations. Further to estimate the cost and drilling resources it is necessary to predict the water table and number of days. For the prediction of the water table depth in different regions, we proposed an ensemble model by using the bagging and boosting method as a candidate learner. The MAE, MSE, and RMSE of the E-WTD model are 3.02,28.65, and 5.35 respectively. In the third module, we present an ensemble model by selecting a decision tree, bagging, and boosting as a candidate learner. The MAE, MSE, and RMSE of the E-NOD model are 0.849,1.46 and 1.20 respectively. However, the TabNet model did not produce the expected results in comparison to the Bagging and Boosting method. Extensive empirical investigations revealed that, compared to competing approaches, the models we developed performed better and yielded superior outcomes across all modules. The results of this research will aid the drilling industry and water boards to optimize the resources for sustainable groundwater extraction and management. In this work, we perform soil layer, water table depth, and number of days prediction. But in the future, we may predict the next depth for the next day, soil color, and drilling point. Now we perform experiments on geological features, but in the future, we may use geological features with environmental factors for the depth of water. We may also extract new features from the existing features. To predict the water table depth, days, and soil layer we apply a machine-learning model with bagging and boosting techniques. But in the future other machine learning methods with boosting or bagging techniques may be used.

VI. CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENT

Any correspondence related to this paper should be addressed to Dohyeun Kim.

REFERENCES

[1] J. L. Wescoat Jr., "Books of note–water in crisis: A guide to the world's fresh water resources edited by Peter H. Gleick," *Environment*, vol. 36, no. 4, p. 26, 1994.

[2] M. H. Lo, J. S. Famiglietti, J. T. Reager, M. Rodell, S. Swenson, and W. Y. Wu, "Grace-based estimates of global groundwater depletion," in *Terrestrial Water Cycle and Climate Change: Natural and Human-Induced Impacts*. 2016, pp. 135–146.

[3] K. Khosravi, M. Sartaj, F. T.-C. Tsai, V. P. Singh, N. Kazakis, A. M. Melesse, I. Prakash, D. Tien Bui, and B. T. Pham, "A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment," *Sci. Total Environ.*, vol. 642, pp. 1032–1049, Nov. 2018.

[4] R. C. M. Nobre, O. C. Rotunno Filho, W. J. Mansur, M. M. M. Nobre, and C. A. N. Cosenza, "Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool," *J. Contaminant Hydrol.*, vol. 94, nos. 3–4, pp. 277–292, Dec. 2007.

[5] H. Tabari, J. Nikbakht, and B. S. Some'e, "Investigation of groundwater level fluctuations in the north of Iran," *Environ. Earth Sci.*, vol. 66, no. 1, pp. 231–243, May 2012.

[6] J. D. Mackay, C. R. Jackson, A. Brookshaw, A. A. Scaife, J. Cook, and R. S. Ward, "Seasonal forecasting of groundwater levels in principal aquifers of the united kingdom," *J. Hydrol.*, vol. 530, pp. 815–828, Nov. 2015.

[7] A. M. Alsalama, J. P. Canlas, and S. H. Gharbi, "An integrated system for drilling real time data analytics," in *Proc. SPE Intell. Energy Int. Conf. Exhib.*, OnePetro, TX, USA, Sep. 2016, pp. 1–11.

[8] H. Wang, C. Ma, and L. Zhou, "A brief review of machine learning and its application," in *Proc. Int. Conf. Inf. Eng. Comput. Sci.*, Dec. 2009, pp. 1–4.

[9] J. Jeong and E. Park, "Comparative application of various machine learning techniques for lithology predictions," *J. Soil Groundwater Environ.*, vol. 21, no. 3, pp. 21–34, Jun. 2016.

[10] B. T. Pham, A. Jaafari, T. V. Phong, D. Mafi-Gholami, M. Amiri, N. V. Tao, V.-H. Duong, and I. Prakash, "Naïve Bayes ensemble models for groundwater potential mapping," *Ecological Informat.*, vol. 64, Sep. 2021, Art. no. 101389.

[11] O. Rahmati, M. Avand, P. Yariyan, J. P. Tiefenbacher, A. Azareh, and D. T. Bui, "Assessment of Gini-, entropy- and ratio-based classification trees for groundwater potential modelling and prediction," *Geocarto Int.*, vol. 37, no. 12, pp. 3397–3415, Jun. 2022.

[12] S. Lee, K.-K. Lee, and H. Yoon, "Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors," *Hydrogeology J.*, vol. 27, no. 2, pp. 567–579, Mar. 2019.

[13] A. A. Nadiri, K. Naderi, R. Khatibi, and M. Gharekhani, "Modelling groundwater level variations by learning from multiple models using fuzzy logic," *Hydrological Sci. J.*, vol. 64, no. 2, pp. 210–226, Jan. 2019.

[14] M. Zare and M. Koch, "Groundwater level fluctuations simulation and prediction by ANFIS- and hybrid wavelet-ANFIS/Fuzzy C-means (FCM) clustering models: Application to the miandarband plain," *J. Hydro-Environ. Res.*, vol. 18, pp. 63–76, Feb. 2018.

[15] J. Zhang, Y. Zhu, X. Zhang, M. Ye, and J. Yang, "Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas," *J. Hydrol.*, vol. 561, pp. 918–929, Jun. 2018.

[16] Y. Ben, C. James, and D. Cao, "Development and application of a real-time drilling state classification algorithm with machine learning," in *Proc. 7th Unconventional Resour. Technol. Conf.*, OnePetro, TX, USA, 2019, pp. 1–12.

[17] F. Sajedi-Hosseini, A. Malekian, B. Choubin, O. Rahmati, S. Cipullo, F. Coulon, and B. Pradhan, "A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination," *Sci. Total Environ.*, vol. 644, pp. 954–962, Dec. 2018.

[18] D. Castiñeira, R. Toronyi, and N. Saleri, "Machine learning and natural language processing for automated analysis of drilling and completion data," in *Proc. SPE Kingdom Saudi Arabia Annu. Tech. Symp. Exhib.*, OnePetro, TX, USA, Apr. 2018, pp. 1–15.

[19] M. Demirci, F. Unes, Y. Z. Kaya, M. Mamak, B. Tasar, and E. Ispir, "Estimation of groundwater level using artificial neural networks: A case study of Hatay-Turkey," in *Proc. 10th Int. Conf. Environ. Eng.*, Aug. 2017, pp. 1–12.

[20] Y. Z. Kaya, F. Üneş, M. Demirci, B. Taşar, and H. Varçin, "Groundwater level prediction using artificial neural network and M5 tree models," *Aerul şi Apa, Componente ale Mediului*, pp. 195–201, 2018.

[21] B. D. Bowes, J. M. Sadler, M. M. Morsy, M. Behl, and J. L. Goodall, "Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks," *Water*, vol. 11, no. 5, p. 1098, May 2019.

[22] O. Kombo, S. Kumaran, Y. Sheikh, A. Bovim, and K. Jayavel, "Long-term groundwater level prediction model based on hybrid KNN-RF technique," *Hydrology*, vol. 7, no. 3, p. 59, Aug. 2020.

[23] B. Yadav, S. Ch, S. Mathur, and J. Adamowski, "Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction," *J. Water Land Develop.*, vol. 32, no. 1, pp. 103–112, Mar. 2017.

[24] P. T. Nguyen, D. H. Ha, M. Avand, A. Jaafari, H. D. Nguyen, N. Al-Ansari, T. Van Phong, R. Sharma, R. Kumar, H. V. Le, L. S. Ho, I. Prakash, and B. T. Pham, "Soft computing ensemble models based on logistic regression for groundwater potential mapping," *Appl. Sci.*, vol. 10, no. 7, p. 2469, Apr. 2020.
[25] W. Chen, X. Zhao, P. Tsangaratos, H. Shahabi, I. Ilia, W. Xue, X. Wang, and B. B. Ahmad, "Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping," *J. Hydrol.*, vol. 583, Apr. 2020, Art. no. 124602.
[26] S. Miraki, S. H. Zanganeh, K. Chapi, V. P. Singh, A. Shirzadi, H. Shahabi, and B. T. Pham, "Mapping groundwater potential using a novel hybrid intelligence approach," *Water Resour. Manage.*, vol. 33, no. 1, pp. 281–302, Jan. 2019.
[27] A. Mosavi, F. S. Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and E. R. Sardooi, "Ensemble boosting and bagging based machine learning models for groundwater potential prediction," *Water Resour. Manage.*, vol. 35, no. 1, pp. 23–37, Jan. 2021.
[28] J. Brédy, J. Gallichand, P. Celicourt, and S. J. Gumiere, "Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches," *Agricult. Water Manage.*, vol. 233, Apr. 2020, Art. no. 106090.
[29] A.-A. Jahanara and S. R. Khodashenas, "Prediction of ground water table using NF-GMDH based evolutionary algorithms," *KSCE J. Civil Eng.*, vol. 23, no. 12, pp. 5235–5243, Dec. 2019.
[30] S. Pradhan, S. Kumar, Y. Kumar, and H. C. Sharma, "Assessment of groundwater utilization status and prediction of water table depth using different heuristic models in an Indian interbasin," *Soft Comput.*, vol. 23, no. 20, pp. 10261–10285, Oct. 2019.
[31] K. Pandey, S. Kumar, A. Malik, and A. Kuriqi, "Artificial neural network optimized with a genetic algorithm for seasonal groundwater table depth prediction in Uttar Pradesh, India," *Sustainability*, vol. 12, no. 21, p. 8932, Oct. 2020.
[32] T. Zhou, F. Wang, and Z. Yang, "Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction," *Water*, vol. 9, no. 10, p. 781, Oct. 2017.
[33] M. A. Mojid, M. F. Parvez, M. Mainuddin, and G. Hodgson, "Water table trend—A sustainability status of groundwater development in north-west Bangladesh," *Water*, vol. 11, no. 6, p. 1182, 2019.
[34] T. Zhao, Y. Zhu, M. Ye, W. Mao, X. Zhang, J. Yang, and J. Wu, "Machine-learning methods for water table depth prediction in seasonal freezing-thawing areas," *Groundwater*, vol. 58, no. 3, pp. 419–431, May 2020.
[35] E. A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, and M. Vaccari, "Groundwater prediction using machine-learning tools," *Algorithms*, vol. 13, no. 11, p. 300, Nov. 2020.
[36] M. Sabah, M. Talebkeikhah, D. A. Wood, R. Khosravanian, M. Anemangely, and A. Younesi, "A machine learning approach to predict drilling rate using petrophysical and mud logging data," *Earth Sci. Informat.*, vol. 12, no. 3, pp. 319–339, Sep. 2019.
[37] H. Fattahi and H. Bazdar, "Applying improved artificial neural network models to evaluate drilling rate index," *Tunnelling Underground Space Technol.*, vol. 70, pp. 114–124, Nov. 2017.
[38] A. Al-AbdulJabbar, S. Elkatatny, A. A. Mahmoud, T. Moussa, D. Al-Shehri, M. Abughaban, and A. Al-Yami, "Prediction of the rate of penetration while drilling horizontal carbonate reservoirs using the self-adaptive artificial neural networks technique," *Sustainability*, vol. 12, no. 4, p. 1376, Feb. 2020.
[39] Y. Zhao, A. Noorbakhsh, M. Koopialipoor, A. Azizi, and M. M. Tahir, "A new methodology for optimization and prediction of rate of penetration during drilling operations," *Eng. Comput.*, vol. 36, no. 2, pp. 587–595, Apr. 2020.
[40] J. Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery, 2020.
[41] A. Puri and M. Gupta, "Review on missing value imputation techniques in data mining," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 2, no. 7, pp. 35–40, 2017.
[42] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, Aug. 2019.
[43] U. Khurana, H. Samulowitz, and D. Turaga, "Feature engineering for predictive modeling using reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–15.
[44] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1560–1571, 2019.
[45] S. Shekarforoush, R. Green, and R. Dyer, "Classifying commit messages: A case study in resampling techniques," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1273–1280.
[46] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.
[47] A. Quemy, "Data pipeline selection and optimization," in *Proc. DOLAP*, 2019, pp. 1–6.
[48] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 101–121.
[49] S. O. Arík and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 6679–6687.
[50] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang, "AutoInt: Automatic feature interaction learning via self-attentive neural networks," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1161–1170.
[51] A. B. Lubis and M. Lubis, "Optimization of distance formula in K-nearest neighbor method," *Bull. Elect. Eng. Inform.*, vol. 9, no. 1, pp. 326–338, 2020.
[52] A. Maheshwari, B. Mehraj, M. S. Khan, and M. S. Idrisi, "An optimized weighted voting based ensemble model for DDoS attack detection and mitigation in SDN environment," *Microprocessors Microsystems*, vol. 89, Mar. 2022, Art. no. 104412.
[53] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of guided regularized random forest for classification and regression tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 88, Jun. 2020, Art. no. 102051.
[54] A. V. Konstantinov and L. V. Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106993.
[55] A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, Jun. 2021.
[56] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30071–30078, Dec. 2020.
[57] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021.

QAZI WAQAS KHAN received the B.Sc. degree in statistics and computer science from the University of the Punjab, Lahore, Pakistan, in 2017, and the M.C.S. and M.S. degrees in computer science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan, in 2019 and 2022, respectively. He is currently pursing the Ph.D. degree in computer engineering with Jeju National University, South Korea. His current research interests include machine learning, federated learning, data mining, natural language processing, computer vision, and the Internet of Things. He was awarded the Merit Fee Waiver Scholarship for the M.S.C.S. degree. He was also awarded the Fee Waiver PEEF Scholarship for the B.S.C. and M.C.S. degrees.

BONG WAN KIM received the B.S. degree in electronics engineering from Hanyang University, Republic of Korea, in 1992, and the M.S. degree in electrical engineering and the Ph.D. degree in computer science and electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1994 and 2000, respectively. He is currently a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI), working on the development of edge computing systems for solving urban problems. His current research interests include edge computing, wireless sensor networks, and data analysis based on artificial intelligence.

RASHID AHMED 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, Republic of Korea, in 2015. Since 2016, he has been with COMSATS University Islamabad, Attock Campus, Pakistan, where he is currently an Assistant Professor with the Department of Computer Science. His current research interests include the application of prediction and optimization algorithms to build IoT-based solutions, machine learning, data mining, and related applications.

ATIF RIZWAN received the B.Sc. degree from the University of the Punjab, Lahore, Punjab, Pakistan, in 2015, and the M.C.S. and M.S. degrees in computer science from COMSATS University Islamabad, Attock Campus, Punjab, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Jeju National University, Republic of Korea. He has good industry experience in mobile and web application development and testing. His current research interests include applied machine learning, data and web mining, analysis and optimization of core algorithms, and the IoT-based applications. He was awarded a fully-funded scholarship for the entire duration of the Ph.D. studies.

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 current research interests include machine learning applications in smart environments, the analysis of prediction and optimization algorithms, big data, and the IoT-based applications

KWANGSOO KIM received the B.S. degree in information engineering and the M.S. degree in computer science from Korea University, Republic of Korea, in 1993 and 1995, respectively, and the Ph.D. degree in computer engineering from Chungnam National University, South Korea, in 2016. Since 1995, he has been a Principal Researcher with the City and Transportation ICT Research Department, Electronics and Telecommunications Research Institute (ETRI), South Korea. His current research interests include spatial information, geographic information systems, location-based services, sensor networks, and the IoT platforms.

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. Since 2004, he has been with Jeju National University, Republic of Korea, where he is currently a Professor with the Department of Computer Engineering. From 2008 to 2009, he was a Visiting Researcher with the Queensland University of Technology, Australia. His current research interests include sensor networks, M2M/IOT, energy optimization and prediction, intelligent service, and mobile computing.