IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Boreholes Data Analysis Architecture based on Clustering and Prediction Models for Enhancing Underground Safety Verification

NAEEM IQBAL[1], ATIF RIZWAN[2], ANAM NAWAZ KHAN[3], RASHID AHMAD[4], BONG WAN KIM[5], KWANGSOO KIM[6] AND DO-HYEUN KIM [7,*]

[1,2,3,6]Computer Engineering Department, Jeju National University, Republic of Korea
[4]Department of Computer Science, COMSATS University Islamabad, Attock Campus 43600, Pakistan
[5,6]Electronics and Telecommunications Research Institute (ETRI), Korea;

Corresponding author: DoHyeun Kim (Email: kimdh@jejunu.ac.kr; Tel.: +82-64-754-3658)

This research was supported by Energy Cloud RD 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 20DCRU-B158151-01). Any correspondence related to this paper should be addressed to Dohyeun Kim.

ABSTRACT During the last decade, substantial resources have been invested to exploit massive amounts of boreholes data collected through groundwater extraction. Furthermore, boreholes depth can be considered one of the crucial factors in digging borehole efficiency. Therefore, a new solution is needed to process and analyze boreholes data to monitor digging operations and identify the boreholes shortcomings. This research study presents a boreholes data analysis architecture based on data and predictive analysis models to improve borehole efficiency, underground safety verification, and risk evaluation. The proposed architecture aims to process and analyze borehole data based on different hydrogeological characteristics using data and predictive analytics to enhance underground safety verification and planning of borehole resources. The proposed architecture is developed based on two modules; descriptive data analysis and predictive analysis modules. The descriptive analysis aims to utilize data and clustering analysis techniques to process and extract hidden hydrogeological characteristics from borehole history data. The predictive analysis aims to develop a bi-directional long short-term memory (BD-LSTM) to predict the boreholes depth to minimize the cost and time of the digging operations. Furthermore, different performance measures are utilized to evaluate the performance of the proposed clustering and regression models. Moreover, our proposed BD-LSTM model is evaluated and compared with conventional machine learning (ML) regression models. The R^2 score of the proposed BD-LSTM is 0.989, which indicates that the proposed model accurately and precisely predicts boreholes depth compared to the conventional regression models. The experimental and comparative analysis results reveal the significance and effectiveness of the proposed borehole data analysis architecture. The experimental results will improve underground safety management and the efficiency of boreholes for future wells.

INDEX TERMS Machine Learning, Deep Learning, Boreholes Data. Data and Predictive Analytics, ROP

I. INTRODUCTION

The revolution in industrial development paved the way towards increasing the urban population rapidly. The rapid growth of social life in urban developed an infrastructure to increase the comfort level for the city dwellers. The developed infrastructure provides everything from water supply to underground subways and rail networks. However, several issues arise due to the outdated and aging infrastructure of big cities, such as water and sewage pipes cracked in big cities. Since the 2000s, the number of accidents has increased due to underground water in urban areas in South Korea. The underground water drilling causes ground depression, such as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

40.2% of sewage pipe damage, 18.7% of water pipe damage, 5.3% of poor excavation work [1].

Digging well is a process to access underground resources such as water, gas, to name of few. Digging technologies have brought a breakthrough change since the first commercial well of oil came into existence through percussion drill technology. Drilling methodologies can be categorized as rotatory and percussion based on rock breaking technique involved. Rotatory methods helped in achieving maximum operational efficiency; however, several other factors need consideration too for digging complex wells. Based on the trajectory and characteristics of well, the digging methods fall into vertical, horizontal and directional technologies. Digging or drilling in a horizontal direction (HDD) is a technique which is gaining enormous attention because of its cost-effective and environment-friendly nature, time involved and land use features of the earth [2]. From its conception till now, HDD is well adopted to geology conditions. However, digging well plays a significant role in the fulfillment of industrial development needs, the economy, and replenishing the needs of safe drinking water for the rest of the world. For the purpose of extracting underground resources digging is done to ample the vast water needs of living beings.

Due to the enormous growth of the groundwater digging process, data growth has already outrun the conventional capacity. Geoscience, hydrogeology, and drilling wells, like various science branches, have also made advancements because of immense technological changes in computing power, remote sensing, and ML. The latest borehore digging approaches generate a massive amount of data and meta-data. The exponential growth of data rate, complexity, variety, and quality is overwhelming, which requires efficient data-driven techniques to cope with such data [3]. The use of big data analytics to aid knowledge discovery is very evident in recent times. For example, in [4], the authors proposed a big data-based analysis for water resource management. In [5], authors employed a big data analysis model to map groundwater potential in South Korea. The major challenge in the data associated with digging groundwater is converting such a massive volume of data into data-driven models.

Analyzing big data is a massive source of information, which is not available and needs to be extracted out of big data. An integral part of big data analytics is data mining (DM) that mine data to trace patterns, relationships between input and output variables, grouping similar data points, or forecasting future outcomes to make informed decisions [6]. Data mining is not limited to big data; it has been in use before the inception of big data, for example, clustering [7], regression [8] and classification [9]. Existing methods seem insufficient in analyzing big data due to multi-dimensional data having different data types and formats. To cope with this, ML techniques are used to process and investigate hidden insights and characteristics that help management to devise effective decisions. For example, the most straightforward unsupervised algorithm is k-mean clustering that can efficiently transform large data sets into samples so that multiple machines can process them [10]. Clustering lies under the umbrella of unsupervised learning, also called data exploration for identifying similar patterns in data [11]. Another method for big data analytics is based on ML algorithms composed of learning modules that are proven to be the backbone of the intelligent systems providing a platform for the analysis of complex and dynamic non-linear systems, such as big data of groundwater wells [12]. However, management of groundwater and optimization of digging well process requires estimation or prediction of hydrological parameters such as next borehole depth. Groundwater borehole's depth point prediction has become a very challenging task [2]. Increased water and drilling demand caused depletion of groundwater resources leading towards abrupt changes in in-depth points. Therefore, a reliable and timely prediction of boreholes depth is required for efficient and informed decision-making to enhance the planning and management of groundwater resources.

With proper utilization and data analysis of boreholes data along with usage of mathematical tools can help prediction of boreholes depth. In the recent advancement in ML techniques, it is quite possible to process hidden characteristics of huge data to build intelligent models for making effective decisions [13], [14]. The most widely used methods for depth rate prediction include Artificial Neural Network (ANN) that are efficient are handling complex non-linear patterns of time series boreholes groundwater data [15]. Nowadays, Deep learning (DL) has become one of the revolutionized research topics in many areas, such as computer vision [16], patterns and object recognition's [17], [18], healthcare [19], etc. DL models enable automatic data representation in a training process and produce the most promising results than traditional ML models. DL models can learn representations automatically from supervised and unsupervised data using a multilayered neural network [20]. DL model is developed based on stochastic optimization, such as Long Short Term Memory Recurrent Neural networks (LSTM-RNN). Recurrent Neural networks (RNN) is an effective time-series data model because it comprises looping structures to process sequence data [21]. However, RNN faces difficulties in learning long-data sequences, which causes exploding gradient problem. To overcome this problem, Long Short Term Memory (LSTM) was introduced to enable feedback connections to perform better than RNN [22]. LSTM is an enhanced variant of the RNN family that has self-connecting hidden layers and a gating structure to handle long-sequences time-series data effectively. Therefore, it is a robust DL model for modeling long-sequences time-series data. The existing prediction model results indicates that the LSTM model performs relatively better compared to traditional ML models [23], [24]. However, some shortcomings exist in traditional DL models [25], for instance, manual representation of features, poor generalization, loss of time, low convergence, and local minima, to the name of a few. To address those problems, a BD-LSTM is proposed to combine forward and backward LSTM models to improve the performance of the prediction model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

BD-LSTM is an effective model for long-range sequences data because two LSTM models concurrently trained on the given long-range sequences data, which enable it to produce better generalization and fast convergence DL model.

The core contributions of the proposed research study are followed as:

- The core contribution of the proposed research study is to utilize data and predictive analysis models to cluster borehole data samples into homogeneous groups based on hidden hydrogeological characteristics and predict boreholes depth for enhancing boreholes efficiency and underground safety verification management.
- Descriptive analyses are employed to utilize data and clustering analysis techniques to process and analyze underlying patterns and hidden characteristics of the boreholes data.
- Statistical and time-series analyses are utilized to investigate historical data of boreholes for underlying patterns and trends.
- Optimal number of clusters is determined using the heuristic elbow curve method.
- GA-based k-means clustering algorithm is developed to cluster borehole data samples into homogeneous groups based on different hydrogeological parameters, such as soil color, land layer, stratum layers and boreholes depth.
- Comparative analysis of the proposed GA-assisted k-means clustering is evaluated and compared with state-of-art clustering techniques.
- Developed BD-LSTM to predict boreholes depth to minimize cost and time for enhancing planning and management of borehole resources.
- Evaluated and compared the BD-LSTM model with conventional ML regression models to demonstrate the significance and robustness of the proposed research study.

The rest of the paper is summarized as follow. Section II presents the related works; Section III presents methodology of the proposed approach. Section IV presents data description and preprocessing. In section V, descriptive data analysis are discussed to investigate hidden characteristics of boreholes data. In section VI, experimental and simulation environment are discussed. In section VII, we present the experimental and performance analysis results. Section VIII concludes the paper with possible future direction.

II. LITERATURE REVIEW

In this section, existing approaches related to the improvements of the digging wells process. Different approaches have been proposed by different researchers to improve and optimize digging process. Several researchers performed wide-scale studies that involved borehole locations, optimization, and scheduling of oil field operations. A study presented in [26] proposed a mixed-integer linear programming model (MILP) known as to determine platform locations to shorten the distance between wells and rigs to increase productivity, revenue and decrease cost. Likewise, [27] also proposed a mixed-integer programming model for ascertaining platform location, assigning them wells, and making decisions regarding pipeline planning and facilitates transportation optimization.

In the recent past rapid growth in technology has generated vast amount of data. AI and ML have emerged as a potent tool for acquiring useful insights from borehole data, prediction and making decisions [28]. With increased boreholes process, ML-based approaches for making informed decisions through prediction based on historical boreholes time series data are becoming more important day by day. Furthermore boreholes depth prediction can be critical for future management of groundwater and digging wells process [29]. Commonly used ML methods employed for prediction include ANN [30], support vector regression (SVR) [31] and K-nearest neighbors (KNN) a feature similarity based method [32]. Several researchers performed wide-scale studies that involved borehole locations, prediction, optimization and scheduling of oil field operations. Digging for underground resources accounts for huge budgets; therefore, any method to reduce time will result in a billion-dollar saving. Rate of penetration (ROP) accounts for the time consumed to drill a well; therefore, the optimization of borehole depth is immensely important in the digging wells. Therefore, a study presented in [33] employed a ML method, Random Forest (RF), on vertical wells dataset for prediction and enhancement of borehole depth using the following parameters, rotations per minute, mud flow rate and weight on bit.

Multiview data with high dimensions are an integral part of big data applications; however, clustering such data is a challenging task due to data features having the same relevance. To deal with this issue, the authors proposed an intelligent weighted k-means clustering technique for multiview data with high-dimensional in various big data applications. The authors formed a distance function based on various weights of views and features for determining object clusters [34]. Moreover, global search is employed by the PSO algorithm for the elimination of sensitivity of initially selected cluster centres. Similarly, in [35], a k-means clustering approach is used for meaningful patterns in learning contexts. This approach aided a collection of heterogeneous data related to learners for improving overall learning process. A study [36] presented a robust density-based affine invariant clustering algorithm using a data depth-based clustering approach to group data samples. Data depth is defined as measures of the centrality of a data point of the given dataset. Moreover, it can detect arbitrary data shapes and forms a stable cluster.

Data-driven techniques are widely used to drive strategic decisions based on historical data interpretation and analysis [37], [38]. For instance, in [39], a two-phase ML-based framework was developed based on bi-directional LSTM to monitor vessel traffic intelligently to enhance the vessel trajectory quality. Another study presented in [40] also developed a bi-directional LSTM based peer-to-peer energy trading platform, which aimed to control the day-ahead dis-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

tribution of energy. In [41], the authors developed an adaptive constrained based dynamic time wrapping (AC-DTW) algorithm to overcome the conventional DTW drawbacks. Furthermore, clustering methods also play a vital role in data mining to recognize patterns and trends. In [42], the authors proposed an enhanced density-based spatial clustering (DB-SCAN) algorithm to group spatial points to obtain optimal clusters for identifying trajectory locations. Both [43] and [44] developed a clustering model using the DB-SCAN algorithm to cluster financial data into homogeneous classes.

Clustering analysis is the process of grouping objects based on similarity into homogeneous clusters in such a way that objects in the same group are similar to each other compared to the objects in other groups [45]. Different researchers attempted to utilize clustering techniques to partition boreholes and hydrogeological data based on underlying patterns for effective digging wells. In [46], the authors proposed a novel k-means clustering algorithm for multi-view data employing a learning mechanism for computation of weights corresponding to new features, which are later used for updation of cluster centers member-ship and view weights. A maximum likelihood-based approach was developed based on clustering techniques to improve the efficiency of the boreholes [47]. In another study presented in [48], the authors determined the linear relationship between digging wells and geological parameters of rock and soil. The authors utilized clustering techniques to investigate time-series data to identify the physical and chemical properties, which change over time during the water digging process [49]. The study presented in [50] also applied clustering techniques to cluster hydro-chemical data samples based on chemical similarity into distinct water groups. Multivariate clustering methods were used to cluster water chemistry data into distinct groups [51]. Partitional clustering analysis utilized boreholes data to identify the relationship between hydrogeological, lithology, and geotechnical [52]. In [53], the authors applied clustering techniques to multivariate time-series data to cluster geochemical anomalies based on underlying and hidden characteristics. C-mean and fuzzy c-mean were applied to partition boreholes and hydrogeological data samples into homogeneous groups [54]. Hierarchical clustering technique was used to the valley soil data to cluster hydrogeological data into distinct groups [55]. In [56], the authors proposed a novel clustering algorithm developed for point data through modification in an existing scale-able algorithm for clustering; involving spatial and temporal distance function. Likewise, a trajectory indexing technique is employed for supporting trajectory-related data.

Prediction models have been employed in many areas as well as in digging wells and hydrogeological engineering. Digging well is considered a costly process due to numerous operations in water, oil, and gas development. Therefore, boreholes depth prediction can be used to decrease costs and operational time to improve digging well efficiency. Several studies have been attempted to process and analyze hidden characteristics of boreholes and hydrogeological data to predict boreholes depth rate. For example, in [57], the authors modeled the depth of the water table based on ANN and SVM. For preparing data, discrete wavelet transform is used. Results depicted that SVM achieved higher performance comparative to ANN. In [58], the authors developed an ANN-based prediction model to predict penetration rate. Next, an artificial bee colony (ABC) algorithm was used to optimize parameters that significantly influence the penetration rate. Another similar study presented in [59], the authors suggested a prediction model to predict penetration rate based on boreholes data using the ANN model to reduce digging cost and time for future wells. Both [60] and [61] proposed PSO-assisted ensemble models to predict the rate of penetration and optimize operational parameters of boreholes to improve the drilling resources. An ANFIS-based prediction model was developed to predict boreholes depth rate and compared experimental results to find an accurate ROP-prediction model [61]. In [62], the authors proposed a computational intelligence-based algorithm to predict the boreholes ROP. The presented method evaluated the predictive performance of learning models based on ML algorithms such as ANN, ELM, SVR, and LS-SVR. Experimental results depict that LS-SVR achieved superior prediction performance with RMSE of 18.27 percent than counterpart solutions. Both [63] and [64] presented ANN-based prediction models to predict the boreholes borehole depth rateto minimize operational digging costs for future wells.

Table 1 presents the summary of the existing studies related to clustering and prediction for digging wells efficiency. To the best of the author's knowledge, all the aforementioned studies are either applied clustering techniques to cluster borehole data or applied prediction techniques to predict borehole depth. The proposed study aims to utilize advanced data and predictive analysis techniques to cluster borehole data based on hydrogeological patterns into homogeneous k clusters and predict boreholes depth for minimizing digging operations cost and time. Furthermore, most of the existing studies attempted to utilize hierarchical and fuzzy-based clustering techniques to cluster boreholes samples to improve the digging wells process. In contrast, our work aims to optimize k-means clustering based on an evolutionary algorithm to cluster boreholes samples into homogeneous groups. Moreover, most of the existing studies utilized ANN to predict the borehole depth to minimize the operational cost of digging wells. In contrast, our proposed work aims to utilize BD-LSTM to predict boreholes depth using borehole data to improve drilling efficiency for future well and compare with the traditional ML models to highlight the significance of the BD-LSTM. Therefore, to the best of our knowledge, it is an open area for researchers to utilize advanced clustering and regression techniques to cluster boreholes samples into k distinct groups based on different hydrogeological characteristics and predict the boreholes depth to improve the digging well efficiency, underground risk evaluation, and underground safety verification management.

TABLE 1. Summary of the existing studies

Existing Approach	Objective	Techniques	Pros	Cons
MILP [26]	This study aimed to find borehole locations to minimize the distance between well and rigs to increase productivity and decrease cost.	Mixed integer linear programming	To assess the tendency of the reservoir and a couple with the multiphase flow in pipelines	High computational time
GWL prediction [29]	The GWL prediction model was developed based on ANN and ANFIS using hydrogeological parameters to predict GWL.	ANN and ANFIS	Accurately predict GWL, achieved R^2 score of 0.96	Hold-out method was used, which can lead to poor generalization issue
Rate of penetration (ROP) [65]	The author's utilized machine learning and data analytics to predict the borehole depth rateto increase the efficiency of the digging process.	Random forest (RF)	It can be used to save time and cost of the digging process because it used surface parameters only	High computational cost due to modifying surface parameters using brute force
Pattern recognition method [47]	The presented study aimed to utilize data-driven methods that searches historic data to identify patterns for future well for efficient drilling.	Gaussian mixture modeling	Provide roadmap based on extracted patterns for future well boreholes	Low flexibility and high computational complexity
Hierarchical cluster analysis [50]	The authors used the pattern recognition method of the hierarchical cluster to partition hydrochemical data into water groups.	Hierarchical clustering	Partitioned water samples into homogeneous groups based on extracted patterns	Considered only sedimentary layer to extract characteristics
ssFCM [53]	The main objective was to cluster multivariate soil geochemical anomalies using a semi-supervised fuzzy c-mean clustering method.	Hierarchical and semi supervised fuzzy c-mean	Combine supervised and unsupervised techniques to extract hidden characteristics of the data	High computational complexity
Cluster analysis [55]	Two types of cluster analysis were performed to identify and partition groups of elements having similar geological structures.	Hierarchical and non-hierarchical clustering	Identified and partitioned elements into distinct groups based on similar geological sources	Difficult to determine the number of clusters for large data, and high time complexity
Boreholes depth prediction [57]	Data-driven approaches were used to predict boreholes depth to improve boreholes efficiency for the future well.	ANN and SVM	Accurately predicted boreholes depth using SVM, comparison of SVM with ANN	Lack of time-series prediction algorithms
borehole depth rate prediction [58]	Developed ANN model based on optimization algorithm to predict borehole depth rateto reduce the cost for effective drilling.	ANN and ABC optimization algorithm	Used optimized set of parameters using ABC optimization scheme	Used Hold-out method to validate data samples
borehole depth rate prediction [61]	Trained ANN model using evolutionary techniques to predict ROP to improve digging wells efficiency.	ANN	Identified influential parameters of the drilling process	High computational complexity

III. PROPOSED METHODOLOGY

This section presents a detailed methodology of the proposed architecture. The main objective of the proposed architecture is to utilize data and predictive analysis models for enhancing efficiency and underground safety management of future boreholes digging.

A. PROPOSED ARCHITECTURE

This subsection presents architecture of the proposed approach. Fig. 1 presents architecture diagram of the proposed approach. The proposed architecture consists of five layers: the acquisition of borehole data, preprocessing of borehole data, descriptive data analysis of borehole data based on different hydrogeological patterns, development of ML-based prediction models, planning and city construction management. The proposed study uses a real borehole dataset acquired from Jeju National University (JNU), South Korea. The dataset consists of different attributes, such as borehole ID, x and y coordinates, altitude, starting depth, ending depth, soil color, land layer, to name a few. Data preprocessing is used to preprocess the acquired boreholes data to increase the reliability and effectiveness of the dataset. It is an important step to transform raw data into a reliable format for pattern extraction and other processes. Data analysis uses preprocessed data to investigate hidden insights and patterns of the boreholes data. Different data analysis models are developed, including statistical, time-series, and clustering analysis, to get hidden insights and patterns of borehole data. Data analysis is an effective process to investigate and extract hidden insights and characteristics from the given preprocessed data, which is important for planning digging wells resources for future boreholes.

Clustering analysis is used to cluster borehole data based on soil color, land layer, and boreholes depth into k distinct clusters to identify and understand data structure. The main objective of the clustering model is to develop an advanced clustering approach to cluster borehole data samples into homogeneous groups based on hidden hydrogeological patterns to improve digging wells efficiency and underground safety management. Therefore, a GA-assisted k-mean clustering algorithm is proposed to cluster borehole samples into distinct k groups. Furthermore, a heuristic method is used to calculate the required number of optimal clusters. The proposed clustering model results are evaluated and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

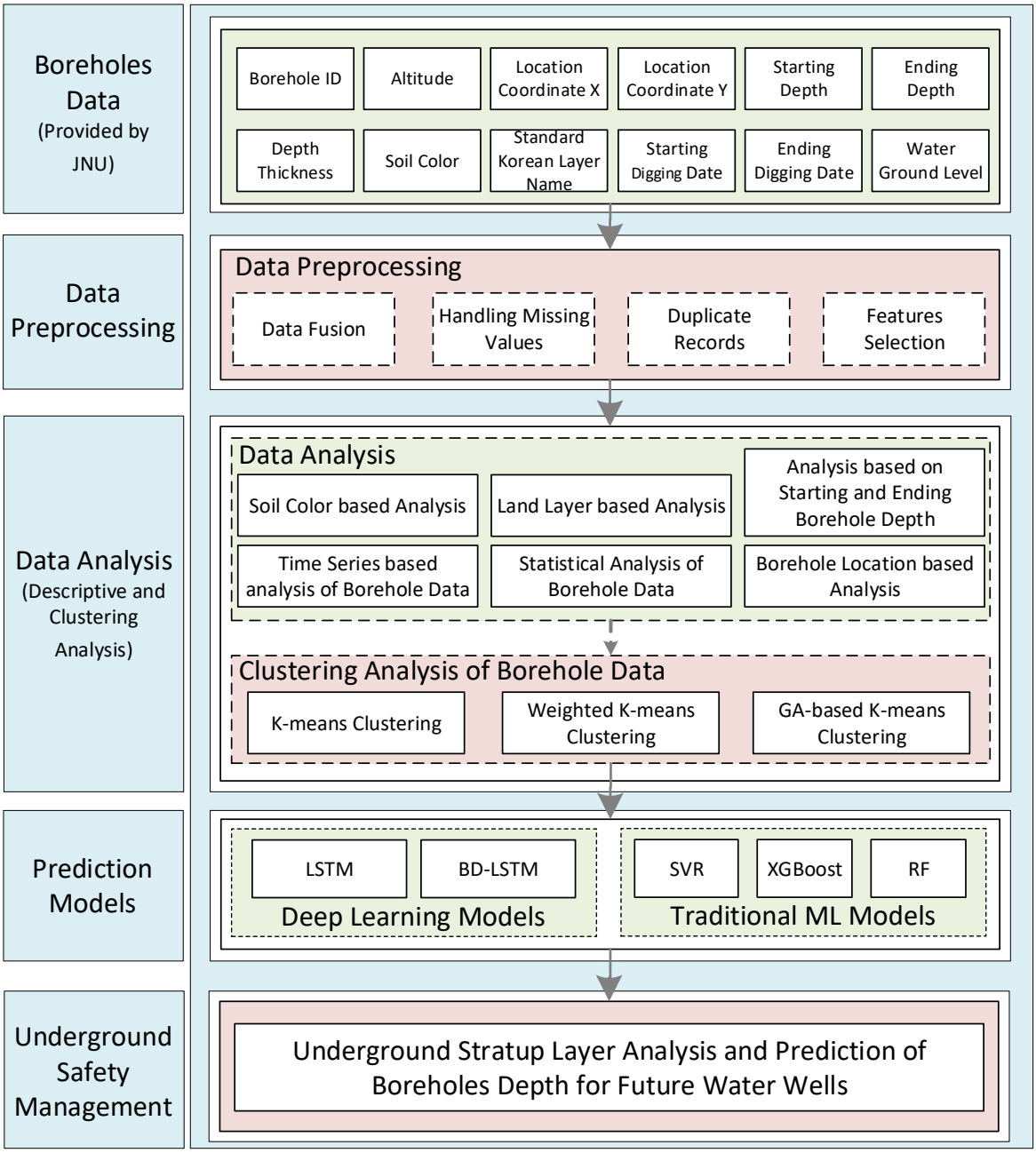


FIGURE 1. Proposed architecture diagram of the borehole data analysis architecture based on data and predictive analysis models.

compared with the traditional k-mean and weighted k-mean algorithms. Next, we develop traditional and deep learning-based prediction models to build an effective model to predict the boreholes depth rate using the history data of boreholes. The main objective of the prediction model is to predict boreholes depth rate to minimize digging operation costs for effective drilling management and planning. The last layer is responsible for effectively planning and managing the drilling resources based on experimental results for enhancing efficiency of underground safety management.

B. PROCESS PROCEDURE OF THE PROPOSED APPROACH

This subsection presents step by step procedure of the proposed architecture. Fig. 2 presents interaction model of the proposed boreholes data analysis architecture. The proposed architecture process consists of different steps: acquisition of boreholes data, preprocessing of boreholes data to get consistent data, data models generation for analysis, ML-based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

prediction models, tuning of hyper-parameters to enhance the performance of the proposed predictive analysis models, execution of ML models to get desired results and storing desired results in repository for future boreholes.

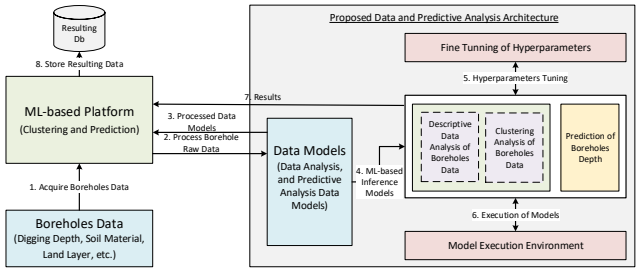


FIGURE 2. Procedure diagram of the proposed architecture.

C. PROPOSED CLUSTERING MODEL

This subsection discusses clustering module, which aims to cluster borehole data based on different features into k distinct groups. In this work, a novel GA-based k-means algorithm using elbow curve method is proposed to cluster borehole data into k distinct groups. Fig. 3 presents proposed optimal k-means clustering mechanism based on elbow curve method. The proposed optimal GA-based k-means clustering mechanism consists of the following components, such as borehole depths data, elbow curve-based computation of optimal k number of clusters, GA-based selection of cluster centroids, clustering of boreholes depths data using k-means clustering technique based on optimal k and optimal centroids values, performance analysis, and clustering results repository. The elbow curve method is a heuristic method, which is used to determine the optimal k number of clusters for the given boreholes input dataset. The genetic algorithm (GA) is used as an optimization scheme, which is used to avoid local optima and discover good initial centroids that lead to superior partitions under k-means. There are different methods used to calculate the distance between data points (borehole data) and centroids, such as Euclidean Distance or Manhattan distance. The k-means clustering uses optimal k and optimal centroids in order to calculate distance, and group the given data instances based on minimum distance. Finally, performance analysis is performed in order to evaluate the performance of the proposed GA-based k-means clustering mechanism. The clustering data is stored in the clustering results repository.

Algorithm 1 presents detail flow of the proposed GA-based k-means clustering approach to cluster borehole data samples into k distinct groups. The proposed algorithm consists of three main modules. First, elbow curve method is used to determine optimal k number of clusters. Second, GA-based optimization scheme is used to determine optimal cluster centroids. Lastly, we used optimal k and optimal clusters centroids to group borehole data into distinct groups to help drilling management.

Algorithm 1: Proposed optimal GA-based k-means clustering using heuristic method.

Input: Borehole Data Samples $D = \{D_1, D_2, D_3, ..., (D_n)\}$, Relevent Features $F = F_1, F_2, F_3, ..., F_n$, k number of optimal clusters, weighted feature W_f

Output: Clustering of Borehole Data samples based on hydrogeological parameters into optimal k clusters

$W_f \leftarrow Borehole_{depth}$
$k \leftarrow ElbowCurveMethod$
for *each iteration* **do**
  $C_{centroids} \leftarrow GeneticAlgorithm(D, k)$
  $R_{clusters} \leftarrow Kmeans(D, k, C_{centroids})$
  $P_{analysis} \leftarrow Evaluation(R_{clusters}, C_{centroids})$
  $StorePerformanceAnalysisResults(R_{clusters})$
  $StoreClusteringResults(R_{clusters})$
end
Function `GeneticAlgorithm(`$Borehole_{data}$, k`)`**:**
  $P_t \leftarrow InitializePopulation$
  $max_{iteration} \leftarrow MaxIterations$
  $M_r \leftarrow MutationRate$
  for $i \in max_{iteration}$ **do**
    for *each individual* $\in P_t$ **do**
      $F_{value}[i] \leftarrow evaluate(individual)$
    end
    $P_{individuals} \leftarrow$Selection(P$_t, F_{value}$)
    $O_{spring} \leftarrow$Crossover(P$_{individuals}$)
    $O_{spring} \leftarrow$Mutation(O$_{spring}$)
    $P_t \leftarrow$MergePopulation(O$_{spring}, P_t$)
  end
  return Centroid set with best fitness value;
Function `Kmeans(`D, k, $C_{centroids}$`)`**:**
  for *each sample* $i \in X$ **do**
    $S_d \leftarrow 0$
    $R_c \leftarrow null$
    for *each cluster centroid* $c \in C_{centroids}$ **do**
      $d \leftarrow EuclideanDistance(i, c)$
      if $d < S_d$ **then**
        $S_d \leftarrow d$; `// Update shortest distance`
        $R_c \leftarrow c$; `// Assign data point` i `to the closest cluster` c
      end
    end
    $C_{label}[i] \leftarrow R_c$; `// Update centroids c using GA`
  end
  return C_{label};

The first step consists of the borehole data acquired from the several organizations. The acquired data is not in a reliable format. First of all, we preprocess acquired boreholes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

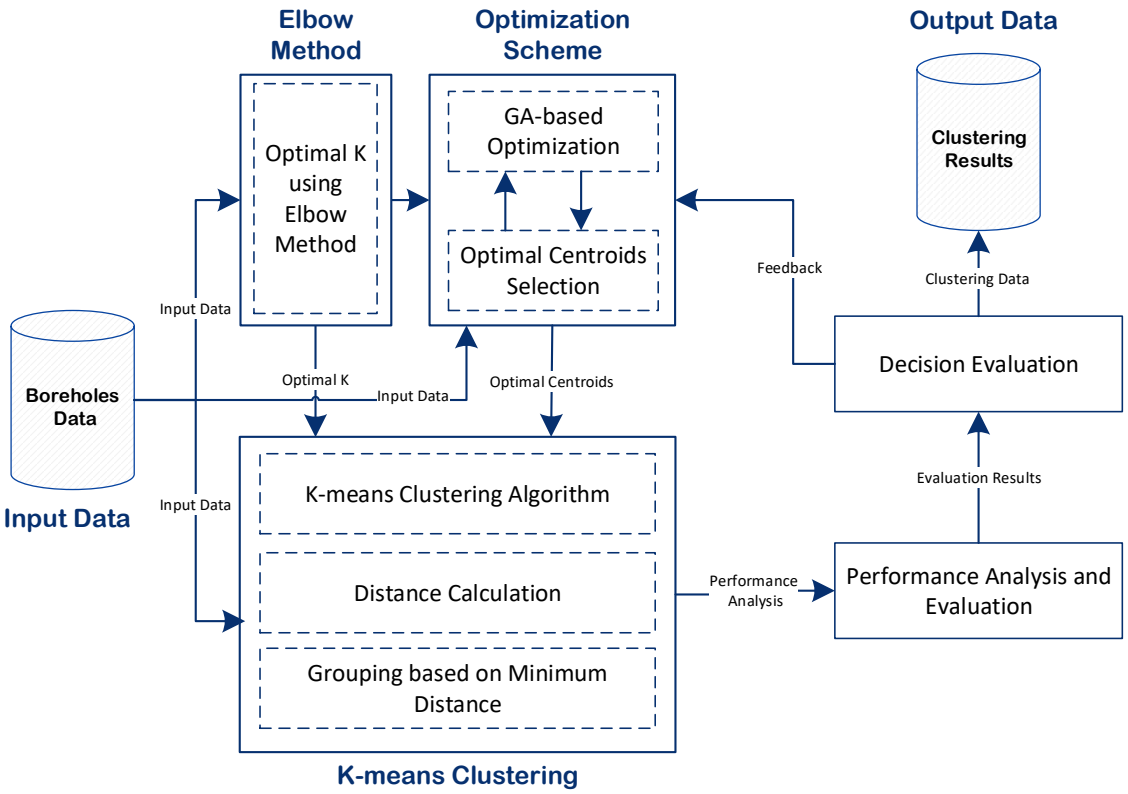


FIGURE 3. Proposed GA-assited k-mean clustering model for clustering borehole data samples.

data to remove irrelevant features and duplicate records to increase the effectiveness of the dataset. The preprocessed data are in reliable form for discovering hidden knowledge and underlying patterns using data mining (DM) techniques.

The k-means clustering is an unsupervised clustering algorithm that requires k (total number of clusters) value in advance to cluster data samples. Therefore, it is still a challenge to calculate the optimal value of k to group data samples using k-means clustering. Therefore, we use an elbow curve method to determine optimal k for the given Borehole Dataset. This is a heuristic method used to determine optimal k. To determine optimal k, an elbow curve method is implemented to run a k-means clustering algorithm on the prepared borehole dataset for a range of k (from 1 to 10) to calculate the sum of squared errors (SSE) for each k. The SSE is defined as shown in equation 1.

$$SSE = \sum_{j=1}^{n_1} |X_{1j} - C_1|^2 + \sum_{j=1}^{n_2} |X_{2j} - C_2|^2 + ... + \sum_{j=1}^{n_k} |X_{kj} - C_k|^2 \quad (1)$$

The SSE can be simplified as follow in equation 2.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |X_{ij} - C_i|^2 \quad (2)$$

Where X_{ij} represents Borehole data sample and C_i represents cluster centroid.

Finally, we visualize SSE for each k using a line chart and select elbow of the curve as an optimal k for the given Borehole Dataset. The main objective is to minimize SSE and select a k with low SSE and elbow of the curve as an optimal number of clusters for the given dataset.

Next, GA-based optimization scheme is used to select optimal centroid values rather than a selection of centroids randomly for k-means clustering. The GA algorithm initializes with a random population, which consists of a set of individuals generated randomly. To generate a new population, different genetic operations are performed, such as selection, crossover, and mutation. Each chromosome (a set of genes) of the population represents k number of centroids. The basic steps are repeatedly applied for several generations to search for appropriate cluster centres in the feature space such that a similarity metric of the resulting clusters is optimized. Thus, we select an optimal number of centroids to cluster boreholes data using k-means clustering technique.

The next step presents k-means clustering, which is a well-known unsupervised algorithm that divides data samples into a k number of sub-groups based on the minimum distance between a data point X_1 and cluster centroid C_i. Also, it is an iterative technique that aims to divide the data samples

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

into predefined k number of distinct non-overlapping clusters where each data sample belongs to only one cluster. It aims to make intra-cluster data samples as similar as possible and keeping clusters as different as possible. It uses distance measure (i.e., Euclidean Distance) to calculate the distance between data points and cluster centroids to assign data points to a particular cluster based on minimum distance. The Euclidean Distance (ED) can be calculated between two points as follow in equation 3.

$$ED_{XY} = \sqrt{\sum_{k=1}^{n}(X_k - Y_k)} \tag{3}$$

In this work, it determines k number of clusters using an elbow curve method in order to divide borehole data samples into k number of clusters. It selects optimize cluster centroids based on the GA mechanism. Thus, it uses optimal clusters centroids values in order to divide borehole data into k clusters by determining a distance between a data point X_i and cluster centroid C_i and assign data sample to a particular cluster based on minimum distance.

D. PROPOSED PREDICTIVE ANALYTICS MODEL

In this section, an architecture of the proposed boreholes depth rate prediction model is presented. Fig. 4 presents the flow of the proposed predictive analytics model to predict the next boreholes depth using boreholes data for enhancing digging wells efficiency by minimizing cost and time. The step-by-step flow of the proposed BD-LSTM is described as follow; acquisition of boreholes data, preparation of boreholes data, normalization of boreholes data, splitting prepared data into training and testing subsets, training proposed BD-LSTM and conventional ML models using training data samples, testing, and evaluation of the learned models. In this study, a min-max normalization method is employed to scale data in a uniform range. The prepared data is split into two subsets with a ratio of 70% (training) and 30% (testing). Different statistical metrics are utilized to evaluate the performance of the proposed BD-LSTM and compare it with conventional ML models. The following statistical metrics are used, such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R^2 score.

Algorithm 2 presents a flow of the proposed predictive analytics model. The training data samples are given as input to the proposed BD-LSTM. First, model parameters are initialized, such as batch size, the total number of training samples, and hyper-parameters, etc. Second, in each iteration, a batch of samples is selected from training data samples D. Each sample of batch D_i is split into sequences to feed into forward and backward LSTM models. Next, experimental results of both forward and backward LSTM models are concatenated using an average mode. The concatenated output is passed to the flatten layer, which is used to convert bi-directional sequences into vector g. Next, vector g is passed to the activation function, such as softmax, to get the desired outcome. Finally, an adaptive learning-based optimization technique is used to tune the hyperparameters to minimize training loss and maximize training accuracy.

Algorithm 2: Proposed predictive analytics algorithm for Borehole depth rate prediction.

Input: Training Data Samples $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)\}$, batch size b, total number of training samples n, model hyper-parameters w

Output: Trained BD-LSTM model

$w \leftarrow initilizeRandomly$
$b \leftarrow batchSize$
$n \leftarrow Size(D)$
for *each iteration* **do**
  for $j \in 1, 2, 3, ..., (\frac{n}{b})$ **do**
    $D_j \leftarrow D$
    $D_j^l \leftarrow splitBatchSamples()$
    $f_{sequences} \leftarrow LSTM_f(D_j^l)$
    $b_{sequences} \leftarrow LSTM_b(D_j^l)$
    $f_{output} \leftarrow f_{sequences}$
    $b_{output} \leftarrow b_{sequences}$
    $c_{sequences} \leftarrow Average(f_{output}, b_{output})$
    $g_{vector} \leftarrow Flatten(c_{sequences})$
    $R_{outcome} \leftarrow activationFunction(g_{vector})$
    $w \leftarrow adamOptimizer$; // Update hyperparameters using adam optimizer technique to minimize training loss
  end
end

IV. DATA PRESENTATION AND PREPROCESSING

This section presents boreholes data acquisition, preprocessing, and descriptive analysis to investigate hidden insights and characteristics of boreholes process.

A. BOREHOLES DEPTH DATA

In this paper, a real boreholes dataset is acquired from several organization. The dataset consists of different attributes, such as borehole ID, x and y coordinates, altitude, starting depth, ending depth, soil color, stratum layer, to name of few. Altitude represents the height of the drilling depth process. Ground Water Level depicts how deep under the ground below the earth's surface is water present. Stratum code represents the layer formed at the earth's surface; it could be sedimentary rock layer soil or igneous rock layer code. The acquired dataset consists of 9,287 data samples along with 12 attributes. There is a 1,987, unique borehole ID (unique borehole locations). Likewise, the geological layer name suggests the name of rock layers or soil at the earth's surface. Soil color defines the composition of soil based on color characteristics. Table 2 presents a summary of the acquired boreholes data with a brief description.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

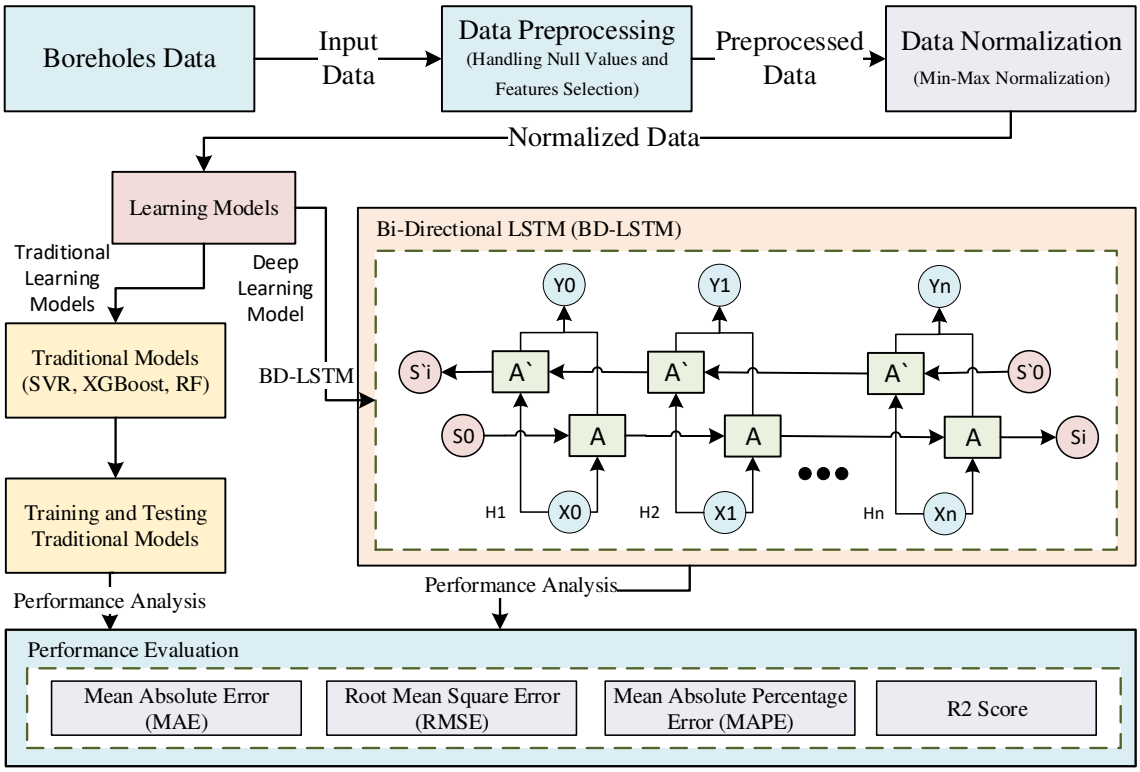


FIGURE 4. Proposed predictive analytics model for borehole depth rate prediction.

B. DATA PREPROCESSING

Data preprocessing is one of the most critical steps in data mining to transform raw data into a reliable format for enhancing the quality and consistency of the data [66]. It helps gain meaningful insights from data by cleaning and organizing it. Using the real dataset, some time leads issue like missing values, noise, and inconsistencies. Therefore data is preprocessed to remove abnormalities. Not dealing with abnormalities later affect the performance of a ML model hence it is an integral part of model building. Data preprocessing comprise of data cleaning, integration, transformation and reduction.

The acquired boreholes data are not in reliable format in order to process and extract hidden hydrogeological patterns and characteristics. Therefore, a solution needed to preprocess dataset in order to handle missing values and duplicate records. In data preprocessing, static and irrelevant data attributes are removed to enhance the effectiveness of the boreholes data. Also, we remove all those data records that don't have values of the following attributes, such as starting and ending boreholes depth.

V. DESCRIPTIVE DATA ANALYSIS OF BOREHOLE DATA

This section presents data analysis of borehole data characteristics to process and analyze hidden insights and underlying patterns from the acquired borehole data. Analyzing data is an easy way to manipulate data, analyze trends, the relationship between independent and dependent variables, and meaningful data patterns and trends. It provides a summary of data involving data interpretation based on analytical and logical reasoning. Different types of data analysis are performed, including statistical and time-series analysis, and clustering analysis to analyze hidden insights of the borehole data.

A. STATISTICAL ANALYSIS

In this subsection, statistical and time-series analyses are performed to investigate borehole data using statistical techniques to seek hidden insights and trends. Fig. 5 is used to analyze the distribution of boreholes data based on rock (land) layer, including Landfill layer, Sedimentary layer, Burlap soil layer, Alluvial soil, Remnant layer, Weathered soil layer, Weathered rock layer, Soft rock layer (ordinary rock), Gyeongam Formation, etc. It can be observed that the majority of borehole data instances belong to a standard layer entitled sedimentary layer.

The sedimentary layer is the most common rock layer formed at the earth's surface due to the accumulation of minerals and organic residues. It is also evident that a small set of borehole data instances belong to the burlap soil layer and residual soil layer. Burlap soil mostly disintegrates with time, while residual soil is formed due to chemical weathering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

TABLE 2. Borehole dataset features and description.

	Feature	Description
1	Borehole ID	It represents a borehole location ID. A unique ID is assigned to each borehole or location.
2	Borehole Resonance	It denotes borehole resonance ID.
3	X	It represents the borehole location coordinate value of X.
4	Y	It represents the borehole location coordinate value of Y.
5	Altitude	It represents an attitude of the digging depth process. High altitude values create a situation where cold weather provisions are needed for proper operation of the digging wells.
6	Ground Water Level	It is used to represents a borehole depth below the earth's surface that is saturated with water.
7	Stratum Code	It represents Korean stratum code.
8	Starting Depth	It represents starting depth of borehole location x for a specific rock unit r. In other words, it is a top depth of a specific layer r for a borehole x.
9	Ending Depth	It represents the ending depth of borehole location x for a specific rock unit r.
10	Geological Layer Name	It is defined as the rock layer that through the borehole passes. It includes the following layers: the landfill layer, sedimentary layer, burlap soil layer, etc.
11	Academic Stratup Name	It represents different academic startup names, such as GM, SM, ET, SP, and CL, etc.
12	Soil Colour	It represents the soil colour of the borehole location.

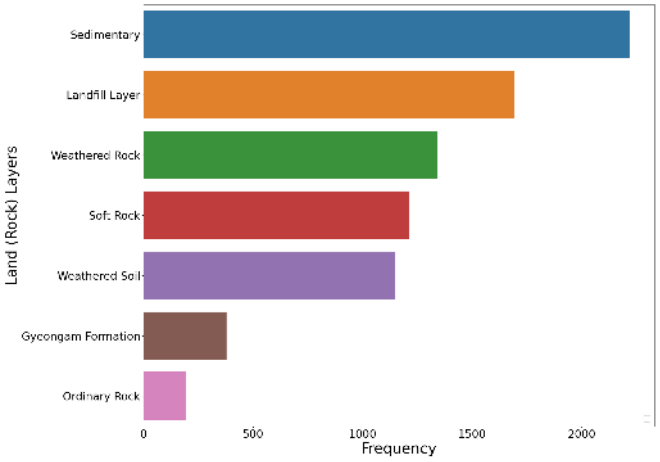


FIGURE 5. Standard Layer based analysis of the prepared Borehole data.

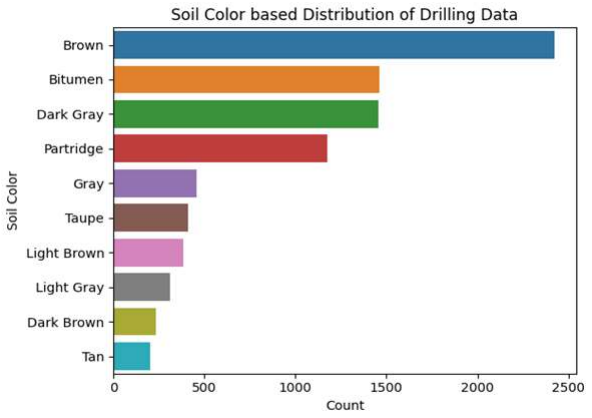


FIGURE 6. Soil colour-based distribution of the prepared borehole data.

residing on top of the parent rock.

Similarly, Fig. 6 presents the frequency distribution of soil colors to investigate the importance of each soil color during the drilling process. The distribution analysis of soil colors shows that a large set of boreholes samples have soil color "Brown" in the selected region to drill the boreholes. In contrast, a small set of boreholes samples have belonged to the soil color "Tan" to drill the boreholes. The soil colors having colors bitumen are the second most occurring distribution compared to other soil colors, such as Dark Gray, Partridge, Gray, Taupe, to name a few.

Furthermore, Fig. 7 shows the distribution of strata layers; academic stratum layers are defined as rock types formed at the earth's surface. The main objective of the frequency distribution analysis of strata layers is to analyze frequency for each stratum layer to drill the boreholes in order to reach the water levels in the selected region. The analysis shows that majority of borehole samples belong to the stratum layer "SM" to drill the boreholes. The second and third highest frequency of strata layers is WR and MR to drill the borehole. It is also evident that the stratum layer is the small frequency to drill the borehole in the selected region.

Furthermore, different features are extracted from the prepared borehole data, including thickness, borehole depth, time spent in terms of days to gain the groundwater level, etc. Thickness is defined as the difference between starting (top) and ending (bottom) borehole depth for a specific rock unit. The basic formula for calculating thickness is defined in equation 4.

$$Thickness(T) = Ending_{depth} - Starting_{depth} \quad (4)$$

Table 3 presents a use case example to calculate thickness for each rock unit in a sequence. It also shows the total borehole thickness as the sum of the thickness of each rock unit.

Similarly, total borehole depth for a borehole is defined as the sum of the difference between starting and ending borehole depth for each rock unit in a sequence. In other words, it is a combination of the thickness of each rock unit for a specific borehole. Equation 5 is used to define total

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

TABLE 3. Use case example to calculate thickness.

Borehole Code	Soil Color	Rock Layer	Starting Depth	Ending Depth	Thickness
B1568BH001	Brown	Landfill Layer	0	3.4	**3.4**
B1568BH001	Brown	Sedimentary Layer	3.4	6.4	**3**
B1568BH001	Partridge	Weathered Soil Layer	3.4	6.4	**2.9**
B1568BH001	Partridge	Weathered Rock Layer	9.3	13.39	**4.09**

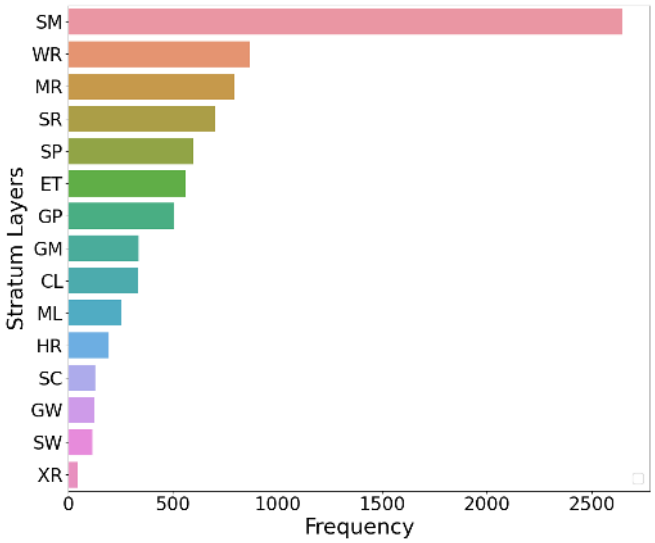


FIGURE 7. Distribution analysis of boreholes data based on stratums layers.

borehole depth.

$$Borehole_{depth} = \sum_{i \in B}^{N} (Ending_{depth}(i) - Starting_{depth}(i)) \tag{5}$$

To simplify further, it can be defined as follows in equation 6.

$$Borehole_{depth} = T_1 + T_2 + T_3 + ... + T_N \tag{6}$$

Next, total time spent on each borehole is calculated as shown in equation 7. In equation 7, $Time_{taken}$ represents time spent on each borehole to gain the water. It is defined as the count of thickness for each specific rock unit of a single borehole. The time resolution for each thickness of a specific borehole is one day. Therefore, the total days spent on each borehole are calculated by counting the total number of thickness combinations for each specific borehole.

$$Time_{taken} = Count\ Combinations\ of\ Thickness\ for\ each\ specific\ borehole \tag{7}$$

In Table 4, a use case example is presented to calculate the time taken (in terms of days) and total borehole depth for the given borehole code (B1568BH001).

Furthermore, different hydrogeological features are extracted from the prepared data, such as the core soil layer

TABLE 4. Use case example to compute $Time_{taken}$ and $Borehole_{depth}$.

Attribute	Value
Borehole Code	B1568BH001
X	204232.7354
Y	540038.4227
Time_{taken}	**4 (days)**
Borehole_{depth}	**13.39 meter**

and core land (rock) layer. Soil color with maximum combinations of thickness is selected as a core soil color for the specific borehole. Similarly, a rock layer with maximum combinations of thickness is selected as the core rock layer for each borehole. Likewise, the stratum layer with maximum combinations is selected as the core stratum layer.

Moreover, statistical analyses are employed to analyze the borehole's depth according to extracted hydrogeological features. Box and whisker plots are considered to evaluate and visualize the extracted features. Box plot analysis is used to divide the borehole data samples to determine a five-number summary, including minimum value, lower median quartile, median, upper median quartile, and maximum value [67]. It is one of the widely used descriptive data analysis methods to summarize a set of borehole data observations on an interval scale. Fig. 8 presents an analysis of extracted core soil colors according to the total combinations of the thickness. The box chart analyzes the distribution of core soil colors according to the thickness combinations to facilitate drilling management for effective resource planning. It is a vital hydrogeological

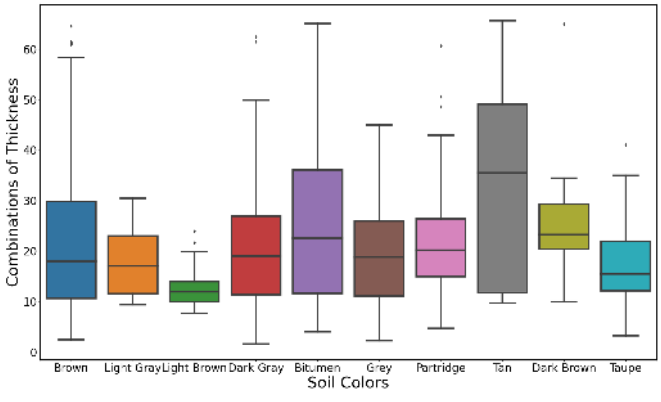


FIGURE 8. Statistical Analysis of boreholes data based on core soil colors.

feature affected by the presence of water due to affecting the oxidation rate. The box-chart analysis shows that core

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

soil color "Tan" has a maximum combination of thickness in the selected region. It can be observed that the maximum combinations of thickness are up to 68 for the core soil color "Tan" in the selected area, which is the highest frequency of thickness combinations than the other core soil colors. It can be observed that the maximum thickness combinations for core soil color "Bitumen" are up to 67, which indicates that the occurrence of the core soil color "Bitumen" is high compared to the rest of core soil colors. Furthermore, Outliers are also identified, as it can be seen that some of the data samples are located outside of the whiskers of the box plot.

Fig. 9 shows a box plot analysis to analyze the distribution of core rock (land) layers according to the combinations of thickness. It aims to analyze the distribution of boreholes based on core rock layers according to the combinations of thickness. As we described earlier, that rock layer with a maximum frequency of thickness is defined as a core rock layer for the specific borehole. The box and whisker plot analysis shows that the length of the land layer "Landfill" is large compared to the other rock layers, which indicates that the maximum frequency of thickness belongs to the Landfill layer to drill a borehole in the selected region. The maximum frequency of the Landfill layer based on the combination of thickness is up to 58, which is the highest frequency compared to the other rock layers. Outliers are also identified that are displayed outside of the whisker of the box plot. Furthermore, the Weathered Soil layer is a minimum median frequency compared to other rock layers.

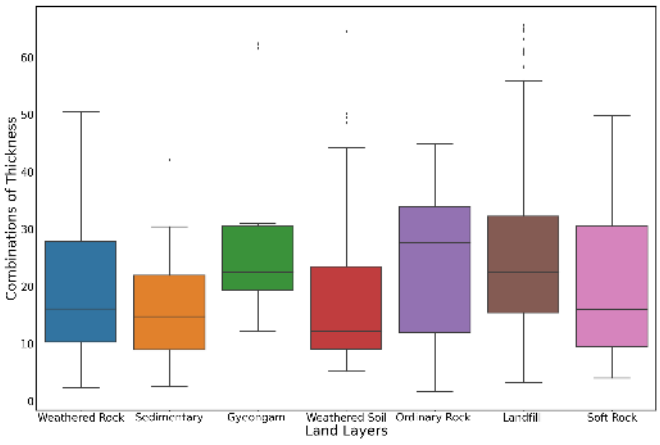


FIGURE 9. Statistical Analysis of boreholes data based on core land layers.

Similarly, Fig. 10 displays the box and whisker plot to analyze the distribution of boreholes based on core strata layers according to combinations of the thickness. As described earlier, thickness is defined as the difference between starting and ending depths for a specific rock unit of the borehole and stratum layer with a maximum combination of thickness as a core stratum layer. The box and whisker plot analysis aims to analyzes the distribution of core strata layers according to the combinations of the thickness (thickness frequency) to get hidden insights of boreholes data. It is evident that the box length of the core SP layer is high, which indicates that most of the data points are scattered to the core stratum layer SP. The maximum analysis shows that the core SP layer is a high frequency of thickness combinations than other strata layers. In contrast, the minimum analysis indicates that the MR is a low combination of thickness in the selected regions than the rest of the strata layers.

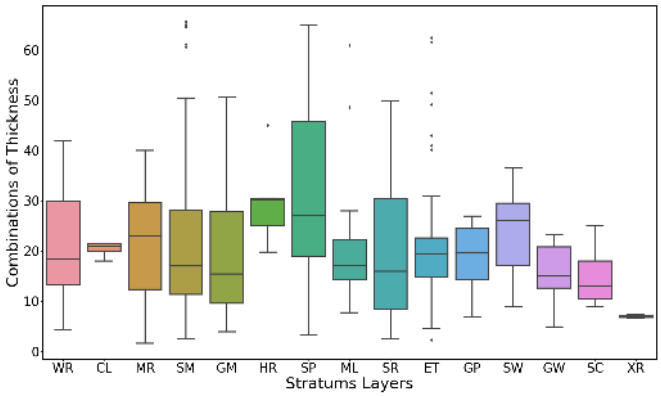


FIGURE 10. Statistical Analysis of boreholes data based on core stratum layers.

B. TIME-SERIES ANALYSIS

Furthermore, time-series analyses are conducted to investigate trends in a particular period. In this work, boreholes data are analyzed based on boreholes depth according to time spent to reach the groundwater levels. Different time-series analyses are conducted to examine the relationship between digging depth and time spent in terms of days. Fig. 11 depicts the relationship between borehole depth and time spent in terms of days to gain the water level. The relationship between borehole depths and time spent in terms of days varies due to the different combinations of the thickness of rock units. It is evident that the total time spent on each borehole fluctuates between 1 to 13 days to gain the water level. Similarly, it can also be seen that the total depth for boreholes varies 74.28 meters. Furthermore, the average time spent on each borehole is approximately 5 days to gain the water level. It can also be observed that the total time spent varies for each borehole due to variations in hydrogeological parameters, such as soil, rock, and land types.

Fig. 12 presents a comparison of drilling depth and groundwater level. The relationship between drilling depth and groundwater level varies due to variation in rock layers. It can be observed that the groundwater level values fluctuate between 0.18 m and 45.47 m. In contrast, borehole depth is up to 74.28 m to gain the water level. It can also be analyzed that the average borehole depth in the selected region is approximately 20.03 m to gain the groundwater level. It can also be interpreted that the average groundwater level in the selected area is 6.73 m.

Furthermore, Fig. 13 depicts the box and whisker plot to analyze drilling depth according to the time taken in terms

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

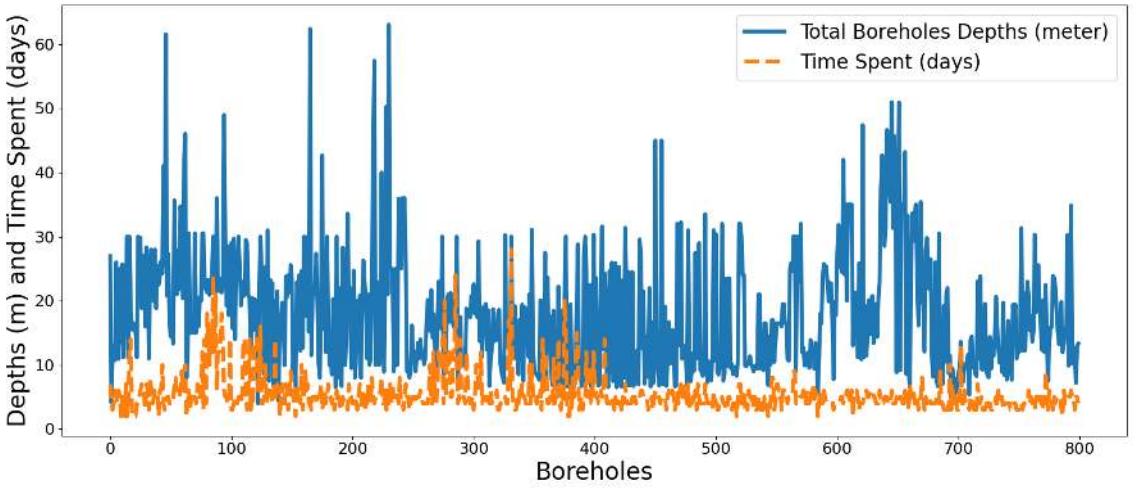


FIGURE 11. Boreholes data analysis based on total depths and time spent.

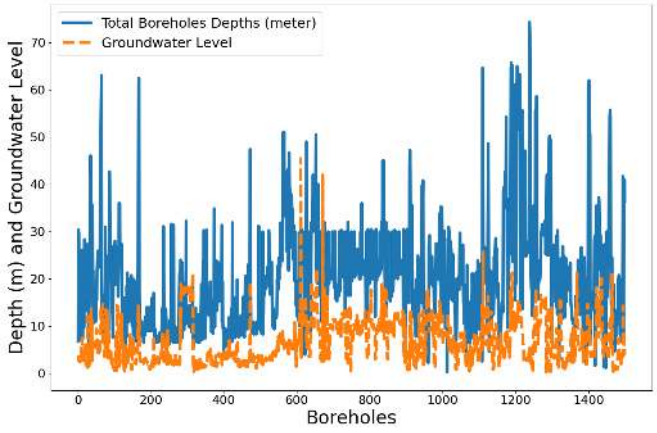


FIGURE 12. Boreholes data analysis based on drilling depth and groundwater level.

of days. It can be seen that the drilling time increases as the borehole depth increases. It can be analyzed that the drilling time varies due to the softness and hardness of the rock units. For instance, it can be interpreted that 2 to 3 days are required to drill up to 24 m. Similarly, 7 to 8 days are required to drill up to 30 m. Furthermore, 11 to 13 days are needed to drill up to 38 m, which indicates the hardness of the rock units compared to other drilling time phases.

Similarly, Fig. 14 presents the box plot analysis to investigate groundwater level according to the drilling time. The analysis shows that the relationship between groundwater level and drilling time varies in the selected region due to the structure of rock layers and other hydrogeological parameters. It can be shown that 1 to 2 days are required for borehole drilling to gain the groundwater level between 2 m to 9 m. It can also be observed that 11 to 13 days are spent to gain the groundwater level up to 7 m, which described that the hardness of rock layers ultimately minimize groundwater level.

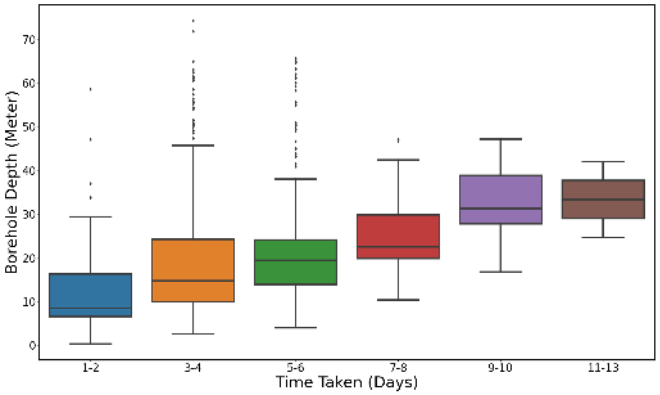


FIGURE 13. Borehole depth analysis according to time taken.

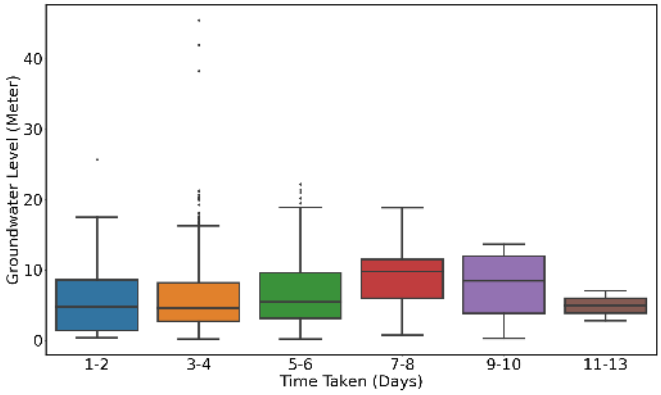


FIGURE 14. Groundwater level analysis according to time taken.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

C. CLUSTERING ANALYSIS RESULTS AND PERFORMANCE ANALYSIS

This subsection presents clustering analysis results of the implemented clustering algorithms. The main objective is to cluster borehole samples into homogeneous groups based on hydrogeological characteristics to increase digging wells efficiency. Fig. 15 shows an elbow curve analysis using a simple k-means clustering algorithm to determine optimal k.

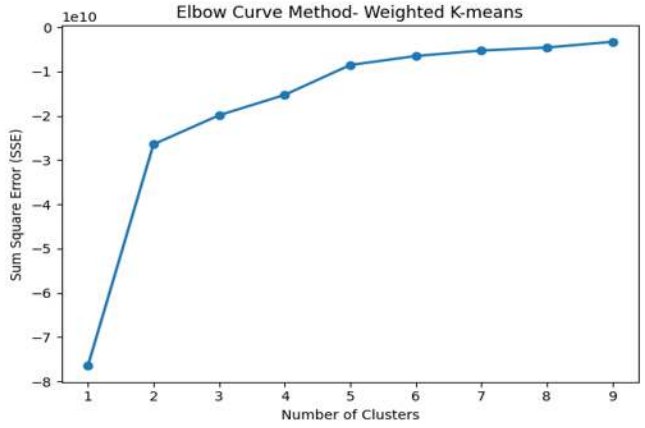


FIGURE 15. Optimal number of clusters using un-weighted k-means clustering.

Fig. 16 presents clustering analysis results obtained using a simple k-mean clustering algorithm. It can be observed that the boreholes samples are clustered into two distinct groups. In this use case example, borehole samples are clustered based on borehole location coordinates to compute the distance between each borehole sample and cluster centroids and assign the borehole sample to the closest cluster. The clustering analysis results are visualized by plotting the data points in their respective clusters.

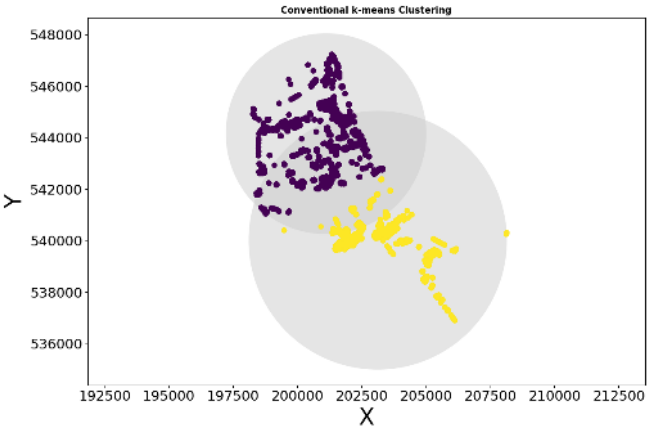


FIGURE 16. Clustering of borehole data based on distance using k-means.

Fig. 17 depicts elbow curve analysis based on proposed features to find an optimal number of clusters. In Fig. 17a, we used borehole location coordinates and borehole depth as an input parameters to run a weighted k-means algorithm for a range of k ($k = 1, 2, ..., 10$) to determine SSE for each k. Based on SSE visualization for each k, it can be observed that elbow of the curve formed at 3; therefore, optimal k is 3 for the given feature set to cluster boreholes data in 3 homogeneous groups. Similarly, in Fig. 17b, the digging depth of the borehole is used as an input parameter to determine an optimal number of clusters. The line chart shows that the elbow of the curve is formed at 2; therefore, optimal k for the given borehole data using borehole depth is 2. Likewise, Fig. 17a and Fig. 17b, Fig. 17c considered soil color as an input feature, and borehole depth is a weighting parameter to determine optimal k. The line chart visualizes the SEE for each k to determine optimal k. It can be seen that the optimal k is 2 to cluster the given borehole data based on hydrogeological patterns. Furthermore, In Fig. 17d, soil color and land layer are considered as the input parameters, and borehole depth is a weighting parameter to select the elbow of the curve as an optimal k. The elbow curve analysis shows that the optimal k is 2 to cluster borehole data based on different hydrogeological characteristics.

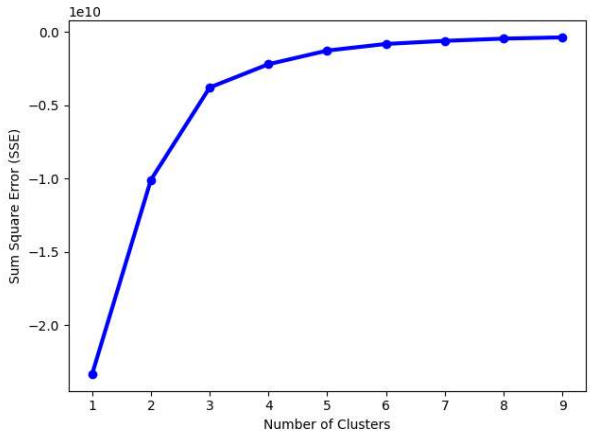
Fig. 18 shows clustering analysis results using GA-assisted k-means clustering algorithms. Borehole location coordinates are considered as the input parameters and the total borehole depth used as a weighting parameter to cluster borehole samples into homogeneous clusters. In this use case example, borehole samples are grouped based on distance using a GA-assisted weighted k-means clustering algorithm. The borehole samples are grouped into 3 distinct groups based on the minimum distance between boreholes and cluster centroids. It can be observed that borehole samples are grouped into 3 distinct clusters, such as cluster 0, cluster 1, and cluster 2.

Fig. 19 presents clustering of borehole samples based on the total borehole depth using GA-assisted k-mean clustering. In this use case, total borehole depth frequency is considered an input parameter to cluster borehole samples based on similar digging frequencies into 2 clusters. In this use case, the soil color attribute is used as a weighting parameter in the clustering process. The borehole samples are grouped into 2 clusters based on the total borehole depths to understand the data structure for effective future boreholes.

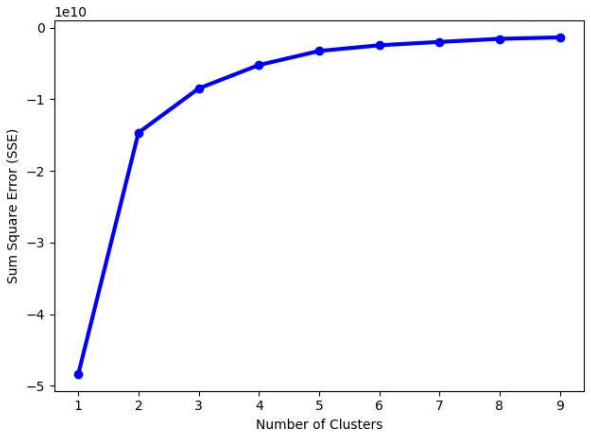
Likewise, in Fig. 20, borehole samples are grouped based on soil color into 2 distinct clusters, such as cluster 0 and cluster 1. The soil color consists of 10 unique colors, including Light Gray, Bitumen, Brown, Dark Brown, Dark Gray, Gray, and Light Brown, etc. In this use case, boreholes location coordinates and soil color are considered as the input parameters and borehole depths as a weighting parameter to cluster borehole samples into 2 clusters to enhance future borehole efficiency.

Similarly, in Fig. 21, borehole data samples are clustered based on two different hydrogeological parameters, such as soil color and land layer. In this use case, location coordinates of the borehole are considered along with soil color, land layer, and total boreholes depths as a weighting parameter

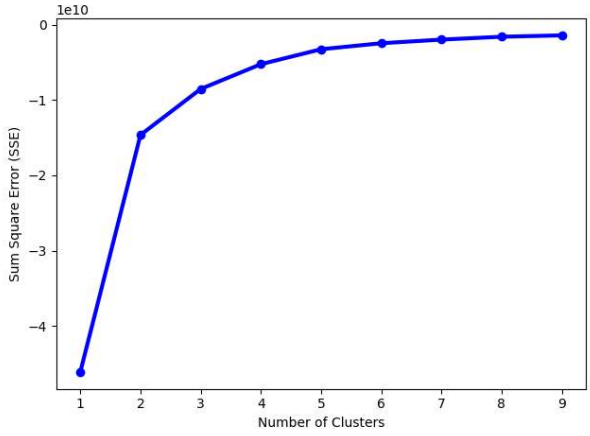
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access



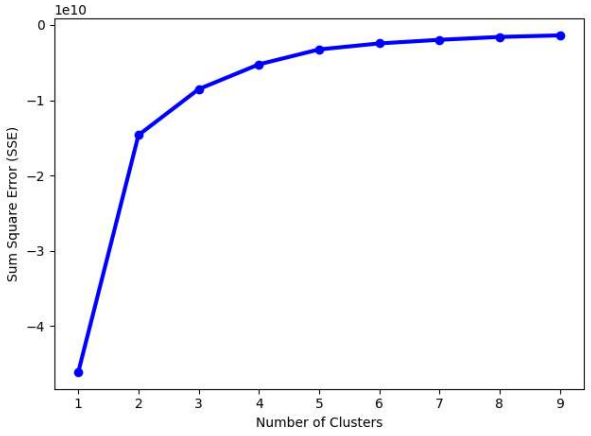
(a) Elbow curve analysis based on borehole location coordinates



(b) Elbow curve analysis based on total borehole depths (total thickness)



(c) Elbow curve analysis based on soil color



(d) Elbow curve analysis based on soil color and land layer

FIGURE 17. Elbow curve analysis based on features set to determine optimal number of clusters.

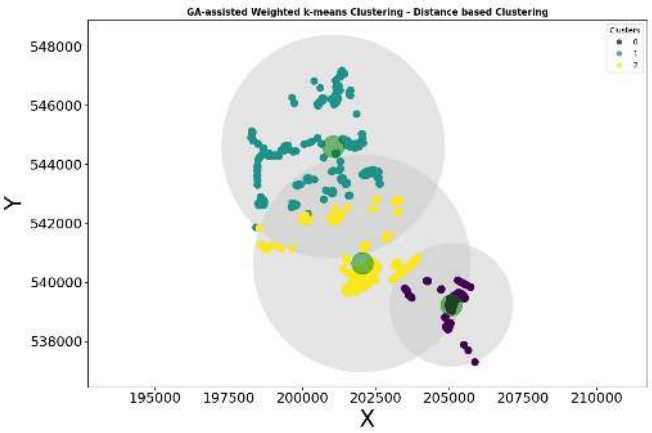


FIGURE 18. Clustering of borehole data based on distance (borehole locations).

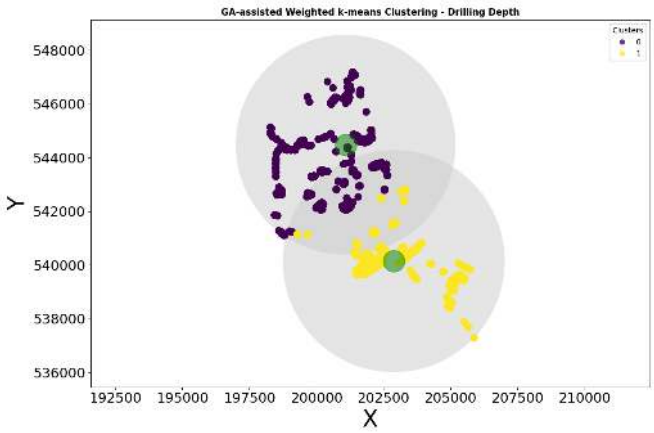


FIGURE 19. Clustering of borehole samples based on digging wells depth.

to group boreholes data samples into 2 distinct clusters. The standard land layer name includes Landfill layer, Sedimentary layer, Burlap soil layer, Alluvial soil, Remnant layer, Weathered soil layer, Weathered rock layer, Gyeongam

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

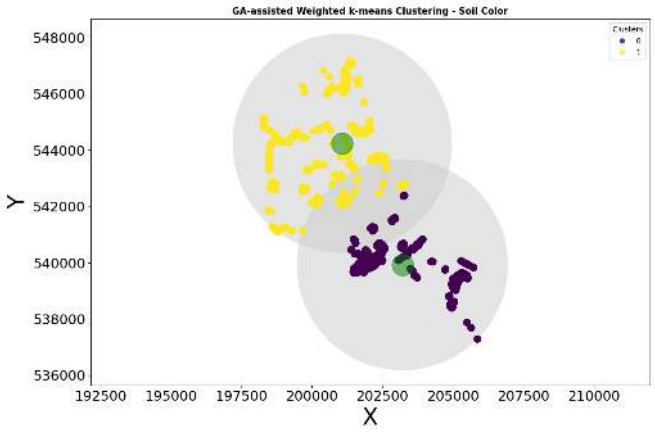


FIGURE 20. Soil color based clustering of borehole samples.

Formation, etc.

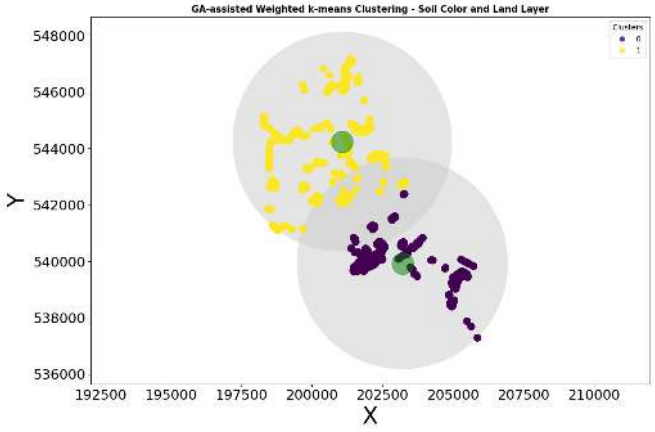
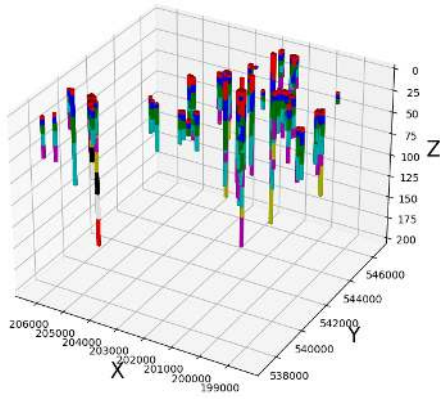


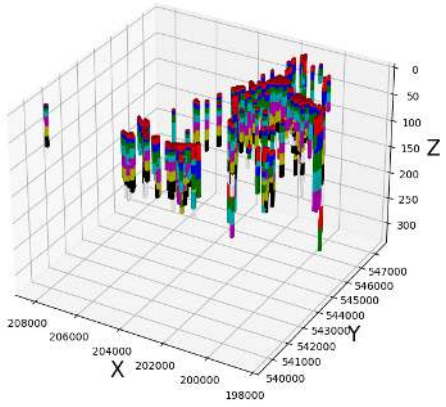
FIGURE 21. Clustering of borehole samples based on soil color and land layer.

Fig. 22 shows the distribution of borehole depths data based on clustering using a distance measure. It shows borehole depths according to the borehole locations in the selected regions. There are three optimal clusters to group borehole samples based on distance into three ($k = 3$) distinct groups, such as cluster label 0, cluster label 1, and cluster label 2. Fig. 22a visualized borehole depths data for cluster label 0 (cluster_label=0). The z-axis represents borehole depths data for the borehole location. It can be observed that boreholes depths data based on the distance for clustering label 0 varies from 13 to 200m. Similarly, Fig. 22b visualized boreholes depths data for cluster label 1 (cluster_label=1). It can be observed that boreholes depths data based on the distance for clustering label 1 varies from 20 to 300m. Likewise, Fig. 22a and Fig. 22b, Fig. 22c show the visualization of total borehole depth for each borehole data sample belongs to cluster 3. It can be seen that the total borehole depth values fluctuate between 10 to 175m.

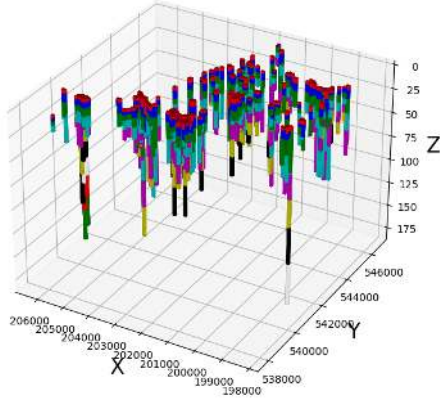
The complete digging process for each borehole consists



(a) Distance based visualization borehole depths (Cluster Label : 0)



(b) Distance based visualization borehole depths (Cluster Label : 1)



(c) Distance based visualization borehole depths (Cluster Label : 2)

FIGURE 22. Comparison of borehole depths based on distance.

of several instances. The minimum number of borehole instances is 3, and the maximum number of borehole instances is 13. Therefore, to visualize total digging depth for each unique borehole, we assign different color codes to borehole instances as shown in Fig. 22. Each color code represents

a borehole depth recorded for a unique set of the following parameters: a land layer, strata layer, and soil color. The total digging depth is calculated by summing the digging depth of each instance (represented by a unique color) for a particular borehole. Furthermore, different evaluation measures are used to evaluate the performance of the clustering algorithms. The following evaluation measures are considered to evaluate and compare the performance of the implemented clustering algorithms, such as Dunn index (DI), Davies–Bouldin index (DBI), Silhouette coefficient (SC), and Calinski–Harabaz Index (CHI).

DI is an appropriate measure to evaluate the performance of clustering algorithms. The main objective of the Dunn index is the identification of compact clusters such that means of clusters lie far apart from one another. The basic formula of the Dunn index is followed in equation 8.

$$DI = \min_{1\leq i\leq m}\left\{\min_{1\leq j\leq n}\left\{\frac{\delta(X_i, X_j)}{\min_{1<k<c}\{\triangle X_k\}}\right\}\right\} \quad (8)$$

DBI is another clustering evaluation measure that evaluates how well clusters are formed based on inherent characteristics of data samples, and lower values depict better clustering performance. It is defined as follows in equation 9.

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{1<i<k}\left\{\frac{\Delta(X_i)+\Delta(X_j)}{\Delta(X_i, X_j)}\right\} \quad (9)$$

SC index is generally used to measure consistency within clusters. It represents how much an object is similar to the rest of its cluster comparative to other clusters. It can be computed as follows in equation 10.

$$SC = \frac{x-y}{max(x,y)} \quad (10)$$

We also used CHI for defining every cluster by a single class. Moreover, it looks for significant variations and differences between expected and observed values. The CHI is defined as follows in equation 11.

$$CHI = \frac{T(B_k)}{T(W_k)} \times \frac{S_n - k}{k-1} \quad (11)$$

Table 5 summarizes the performance evaluation of the proposed clustering algorithms. In this study, borehole data are analyzed and grouped using four different hydrogeological features based on weighted k-mean and GA-assisted k-mean clustering algorithms. Furthermore, different performance measures are employed, such as DI, DBI, SC, and CHI, to evaluate the performance of the implemented clustering algorithms. It is evident that the GA-assisted k-mean clustering algorithm performed well and improved the performance. The performance of the GA-assisted k-mean clustering algorithm in terms of DI for all feature types is 1.092, 1.527, 1.673, and 2.454, which indicates that the DI values of the proposed clustering model higher compared to the weighted k-means clustering algorithm. Furthermore, the estimated value of CHI of the proposed clustering algorithm for all feature groups is 3452.73, 3656.76, 3786.77, and 3987. The CHI analysis shows that the score of the ratio of average between and inter-cluster dispersion of the proposed clustering model is higher than the conventional weighted k-means clustering model. Similarly, in terms of DBI, our proposed model performed relatively better compared to a conventional weighted k-mean clustering algorithm. The DBI score of the proposed GA-assisted k-mean clustering algorithm for all four features set is 0.865, 0.675, 0.721, and 0.454. It is evident that the DBI values of the proposed clustering algorithm are lower than weighted k-means clustering, which indicates that our proposed model performed well in the clustering process.

VI. IMPLEMENTATION ENVIRONMENT OF THE PROPOSED ARCHITECTURE

This section discusses experimentation environment of the proposed approach. The development of proposed models is done using Python programming language. Moreover, preprocessing of data, clustering, and regression analysis are implemented using well-known libraries known as Numpy, Pandas, Scikit-learn, Keras, and TensorFlow. Table 6 presents implementation setup of the proposed approach.

Fig. 23 depicts simulation and implementation flow of the proposed work. The proposed work used Python as a core programming language to implement functionalities of the all modules. The implementation flow of the proposed work includes the following steps; borehole data acquisition, preprocessing of borehole data, prepare borehold data models based on different hydrogeological parameters, such as soil color, land layer, thickness and boreholes depth, to name a few. The proposed work consists of the two core modules; descriptive data analysis, and predictive analysis module to facilitate drilling management for effective planning and management. Descriptive data analysis includes statistical and time-series analysis to investigate underlying patterns of the boreholes data. Finally, experimental results are stored in their respective data repository.

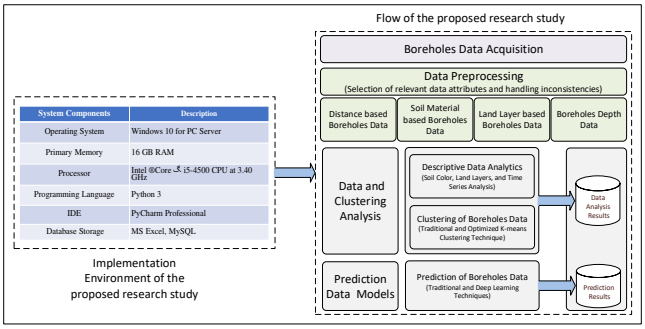


FIGURE 23. Simulation environment of the proposed time-series analysis.

TABLE 5. Performance analysis of the implemented clustering algorithms.

Feature	Model	DI	DBI	SC	CHI
Distance between boreholes	Weighted k-mean	0.524	0.972	0.311	3278.61
	GA-assisted k-mean	1.092	0.882	0.447	3452.73
Boreholes Depth based clustering	Weighted k-mean	0.787	0.792	0.467	3587.98
	GA-assisted k-mean	1.527	0.675	0.498	3656.76
Soil Color based clustering	Weighted k-mean	0.798	0.721	0.517	3631.89
	GA-assisted k-mean	1.673	0.592	0.578	3786.77
Soil Color and Land Layer	Weighted k-mean	1.235	0.678	0.618	3798.75
	GA-assisted k-mean	2.454	0.454	0.765	3987.51

TABLE 6. Implementation setup of the proposed approach.

System Components	Description
Operating System	Microsoft windows 10 (64-bits)
CPU	Intel ®Core ™ i5-4300 CPU at 3.40 GHz
RAM	16 GB
Programming Language	Python
Storage	MySQL
IDE	PyCharm Professional

VII. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

In this section, experiment results of the proposed BD-LSTM and conventional models are discussed. The main objective of the proposed predictive analytics model is to predict boreholes depth accurately to increase digging efficiency for future boreholes. In this study, an advanced prediction model is proposed and compared with conventional regression models, such as SVR, RF, and Extreme Gradient Boosting (XGBoost). Furthermore, prediction results of the proposed BD-LSTM are evaluated and compared with conventional regression models using following evaluation measures, such as MAE, MSE, RMSE, and R^2 score.

Fig. 24 presents comparative analysis of the implemented prediction models in terms of MAE and MSE.

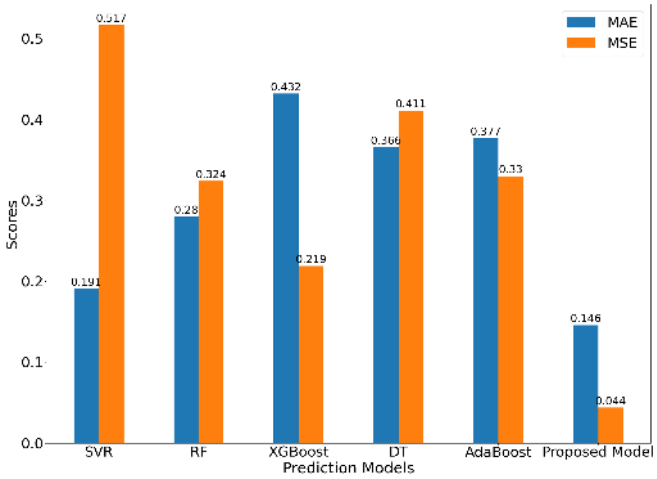


FIGURE 24. Performance evaluation in terms of MAE and MSE.

It depicts that our ensemble prediction model using stacking technique has achieved the highest R^2 score of 0.973. It is also evident that performance of ensemble prediction model using Mean is also up to the mark due to its ability to map temporal correlations and handle long term dependencies. Hence, our proposed ensemble prediction model using stacking technique outperformed all other implemented models and produced relatively better prediction results.

Similarly, Fig. 25 is used to evaluate and compare proposed model with conventional ML models. The R^2 score is used to evaluate and compare the performance of the implemented regression models. The R^2 score of the proposed model is 0.989, which indicates the significance and correctness of the proposed prediction model. The prediction performance of the conventional ML models, such as SVR, RF, and XGBoost, is 0.872, 0.92, and 0.92, respectively. Hence, our proposed BD-LSTM model performed significantly better and outperformed the conventional ML model to predict boreholes depth rate for effective drilling management, planning, and underground safety management.

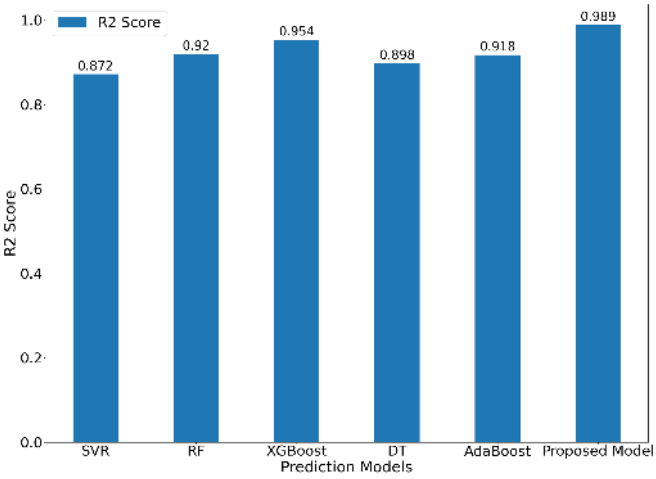
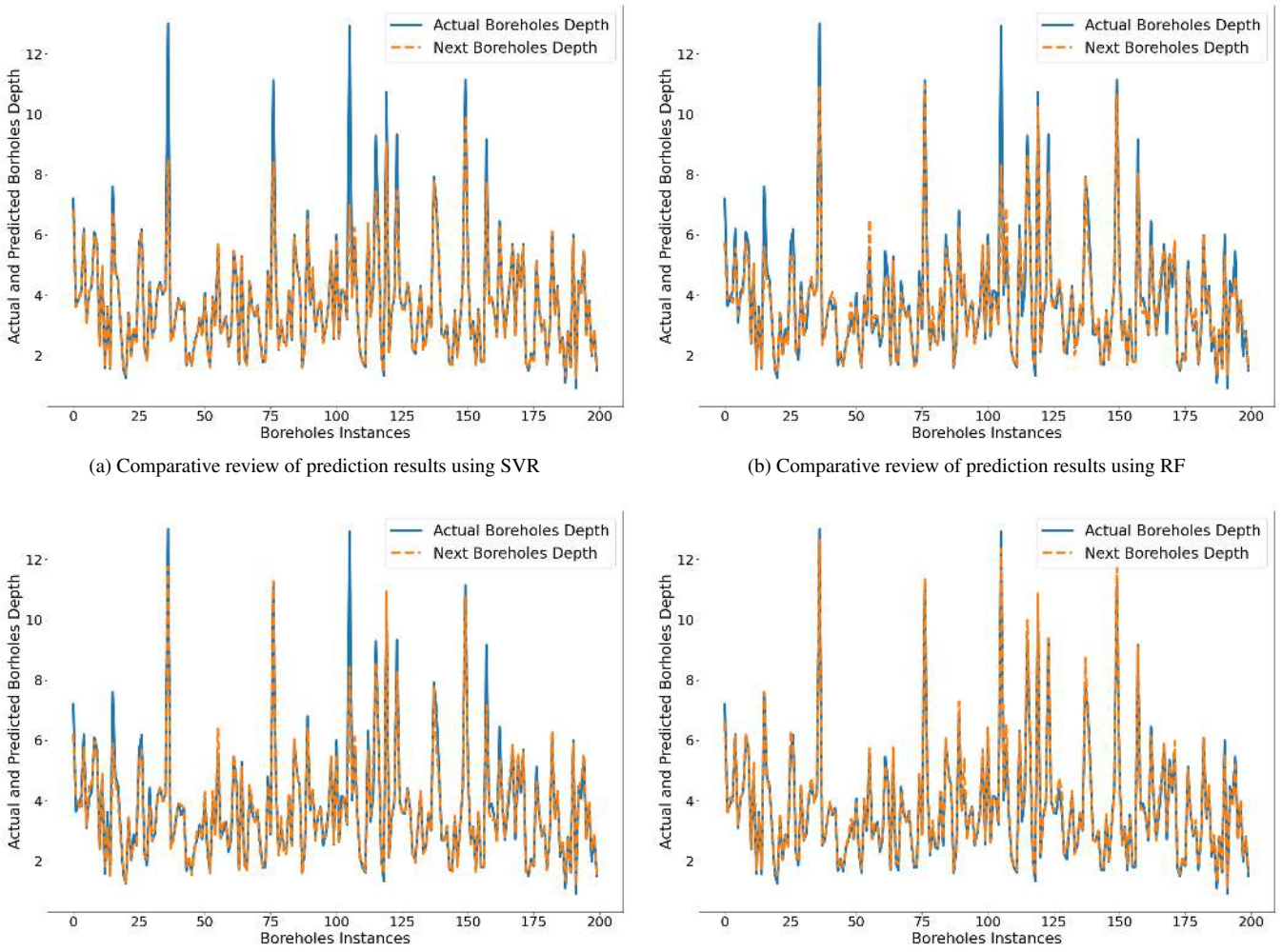


FIGURE 25. Performance evaluation in terms of R^2 Score.

Fig. 26 depicts a comparative review of the observed and predicted boreholes depth using proposed BD-LSTM and other conventional regression models. Fig. 26a presented a comparative review of observed boreholes depth and predicted boreholes depth using conventional SVR. It can be observed that the SVR model produced high prediction errors compared to other implemented regression models. The difference between observed and predicted boreholes depth

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access



(a) Comparative review of prediction results using SVR

(b) Comparative review of prediction results using RF

(c) Comparative review of prediction results using XGBoost

(d) Comparative review of prediction results using proposed model

FIGURE 26. Comparative review of the proposed prediction model with conventional ML regression models for borehole depth rate prediction.

is high; MAE, MSE, and RMSE of the conventional SVR are 0.191, 0.517, and 0.719, respectively. The R^2 score of the conventional SVR is 0.872, which indicates that the conventional SVR performed inaccurately compared to other implemented regression models. In Fig. 26b, a comparative review of prediction results obtained using RF is presented. It is evident that the difference between actual boreholes depth and next boreholes depth is low compared to the conventional SVR. The estimated MAE, MSE, and RMSE of the RF are 0.280, 0.324, and 0.569, respectively, which indicates that the RF performed accurately compared to the conventional SVR. The R^2 score of the RF is 0.92, which reveals that RF outperformed conventional SVR to predict boreholes depth for future wells accurately. Similarly, Fig. 26c shows a comparison of observed and predicted boreholes depth using XGBoost. The XGBoost model performed accurately compared to the RF and SVR; the R^2 score of the XGBoost model is 0.954, which shows that the XGBoost model slightly improves the prediction performance compared to other conventional regression models. Likewise, in Fig. 26d, a comparative evaluation of observed and predicted boreholes depth using proposed BD-LSTM is presented. It is found that the difference between observed and predicted boreholes depth is low compared to the conventional ML regression models, which indicates that the proposed BD-LSTM significantly performed well and improves the prediction accuracy compared to other models. It can be observed that our proposed BD-LSTM performed accurately and significantly as compared to the traditional ML models, such as SVR, RF, and XGBoost. The estimated MAE, MSE, and RMSE are 0.146, 0.044, and 0.210, respectively. Hence, our proposed BD-LSTM performed accurately, improved accuracy, and outperformed the conventional regression models to enhance boreholes efficiency for future wells.

Finally, Fig. 27 presents comparative analysis of actual and predicted borehole depth rate using proposed BD-LSTM and other traditional ML models. It is found that the prediction error of the proposed BD-LSTM is low as compared to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083175, IEEE Access

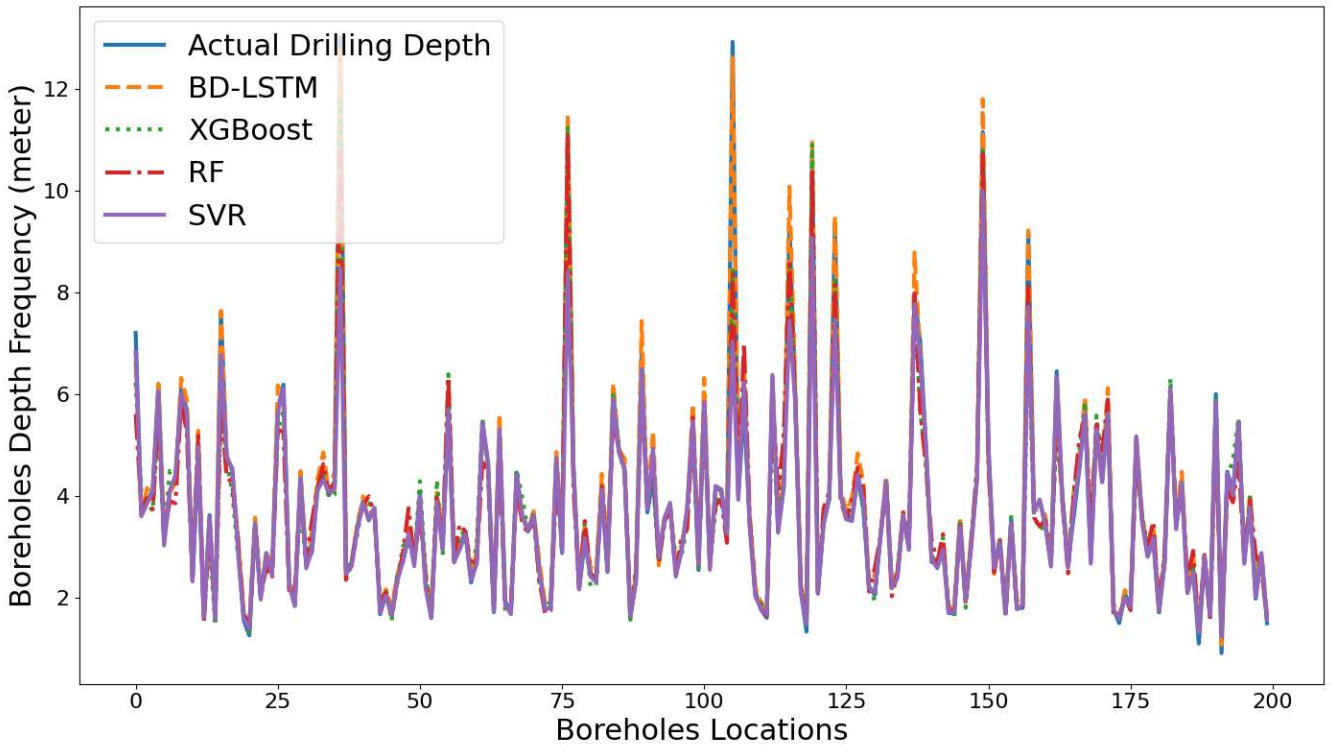


FIGURE 27. Comparative analysis of actual and predicted boreholes depth using proposed BD-LSTM and conventional ML models.

traditional ML-models. Our proposed model performed accurately to predict borehole depth rate for enhancing digging well efficiency by minimizing operational cost for future boreholes. It can be observed that the proposed model significantly improves the prediction performance as compared to other models.

Furthermore, we utilize different statistical metricise, such as MAE, MSE, RMSE, and R^2 score, to evaluate the regression models performance. Accuracy of prediction can be determined using various measures. A low error means a better prediction performance. We can say that accuracy is used to determine the difference between observed output and predicted output. If we want to compare different prediction methods keeping the dataset same every accuracy measure will produce different result and hence different performance. To compare the results with other techniques MAE, MAPE, RMSE, and R-squared (R^2) score and is used for performance evaluation of model [68].

Mean absolute error (MAE) is defined as the average of the absolute differences between predicted and actual values. The absolute error is defined as the difference between predicted and actual value. It is used as a standard to measure error for continuous values. It is defined by equation 12:

$$MAE = \sum_{i=1}^{n} |y_{predicted} - y_{actual}| \tag{12}$$

Mean square error (MSE) is calculated as a mean of square differences between predicted and actual values. MSE is calculated as follows in equation 13.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{predicted} - y_{actual})^2 \tag{13}$$

RMSE is a standard deviation of the prediction errors. It is calculated by taking the square root of the MAE. It is commonly used in regression models to evaluate the forecasting results. RMSE for the continuous variables fluctuates between 0 and ∞, whereas 0 stipulates that the prediction model performed accurately, and a large value indicates that the error between predicted and actual observation is high. It is defined as follow in equation 14.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |(y_{predicted} - y_{actual})|^2}{n}} \tag{14}$$

Another evaluation measure used by this study is R^2. It is used as a determination coefficient to measure the regression model's performance using unseen data similar to MSE. It stipulates the proportion of the variance of predicted and actual variables.

$$R^2\ Score = 1 - \frac{\sum(y_{actual} - y_{predicted})^2}{\sum(y_{actual} - \bar{y}_{predicted})^2} \tag{15}$$

Table 7 presents performance evaluation and comparison of the proposed BD-LSTM with other traditional ML models,

such as SVR, RF, and XGBoost, to name a few. The R^2 score of the proposed BD-LSTM model is 0.989, which indicates the significance and effectiveness of the proposed work. In contrast, the R2 score of the conventional regression models, such as SVR, RF, Lasso (L1), XGBoost, decision tree (DT) regression , AdaBoost, is 0.872, 0.920, 0.81, 0.954, 0.898, 0.918, respectively. The prediction error of the proposed BD-LSTM is low as compared to the traditional ML-models. The estimated MAE and MSE values of our proposed BD-LSTM model are 0.146 and 0.044, respectively. It can be observed that the proposed model significantly improves the prediction performance as compared to other baseline models. Hence, our proposed model accurately predicts borehole depth rate to increase the underground digging efficiency and public safety.

TABLE 7. Performance evaluation of proposed predictive analytics models.

Proposed Models	MAE	MSE	RMSE	R^2 Score
SVR	0.191	0.517	0.719	0.872
RF	0.280	0.324	0.569	0.920
Lasso	0.641	0.766	0.875	0.81
XGBoost	0.432	0.219	0.187	0.954
DT	0.366	0.411	0.641	0.898
AdaBoost	0.377	0.331	0.575	0.918
Proposed Model	**0.146**	**0.044**	**0.210**	**0.989**

VIII. CONCLUSION

Digging well problems possess increasing complexity and uncertainty levels due to hydrogeological variations. However, the application of different mechanisms in borehole data analysis would lead to significant business value and avoid the occurrence of accidents. This research study presented a borehole data analysis architecture based on data and predictive analysis using time series borehole data to improve the digging efficiency, underground risk evaluation, and underground safety management. This study utilized a real boreholes dataset acquired from JNU, South Korea. The main contribution of the proposed study was to employ data and predictive analysis models to discover underlying patterns and predict borehole depth for enhancing the planning and management of borehole resources. Different descriptive data analysis models were employed to investigate historical data of boreholes, such as statistical analysis, time-series analysis, and clustering analysis models. An advanced GA-assisted k-mean clustering algorithm was developed to cluster borehole data samples based on hidden hydrogeological characteristics into homogeneous groups. The main objective of the clustering analysis was to identify hidden patterns to partition boreholes samples into k distinct groups based on four different feature models: distance, borehole depth rate, soil color, land layer, and stratum layer, etc. The clustering approaches were implemented to cluster borehole data samples into distinct groups to identify the history of data patterns for effective resource planning and underground safety management. Furthermore, different evaluation metrics were used to evaluate and compare the proposed clustering approach with the conventional weighted k-means clustering algorithm. The experimental results show that the proposed GA-assisted k-mean clustering technique performed well and relatively better than the conventional weighted k-means clustering algorithm. To predict the borehole depth rate, a BD-LSTM model is developed to predict boreholes depth to minimize the operational and planning cost of the digging for future boreholes. The proposed BD-LSTM performed accurately and outperformed the conventional ML models, such as RF, SVR, L1, XGBoost, DT, and AdaBoost. The results depict that data and predictive analysis models can solve problems that are otherwise impractical or have resulting inaccuracies. The performance evaluation and comparison demonstrate the effectiveness of the proposed model compared to the traditional models. The R^2 score of the proposed BD-LSTM is 0.989, which indicates that the proposed model accurately predicts boreholes depth compared to the conventional ML models. In contrast, R^2 score of the SVR, RF, L1, XGBoost, DT, AdaBoost, is 0.872, 0.920, 0.81, 0.954, 0.898, and 0.918, respectively. Furthermore, our proposed BD-LSTM model significantly performed well in terms of MAE and MSE. The MAE and MSE error of the proposed BD-LSTM is 0.146 and 0.044, which indicates that the difference between observed and predicted boreholes depth is low compared to the conventional model, such as SVR, RF, L1, and XGBoost. The experimental results will improve the overall digging well process, holistic management of groundwater resources, city construction, risk assessment, and underground safety management.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

[1] "Ministry of environment to carry out detailed investigation of old sewage pipes within the year," lhttp://www.waterjournal.co.kr/news/articleView.html?idxno=46623, [Online; accessed on: Feb. 15, 2021].

[2] F. Patino-Ramirez, C. Layhee, and C. Arson, "Horizontal directional drilling (hdd) alignment optimization using ant colony optimization," Tunnelling and Underground Space Technology, vol. 103, p. 103450, 2020.

[3] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International journal of information management, vol. 35, no. 2, pp. 137–144, 2015.

[4] R. Chalh, Z. Bakkoury, D. Ouazar, and M. D. Hasnaoui, "Big data open platform for water resources management," in 2015 International Conference on Cloud Technologies and Applications (CloudTech). IEEE, 2015, pp. 1–8.

[5] S. Lee, Y. Hyun, and M.-J. Lee, "Groundwater potential mapping using data mining models of big data analysis in goyang-si, south korea," Sustainability, vol. 11, no. 6, p. 1678, 2019.

[6] P. Russom et al., "Big data analytics," TDWI best practices report, fourth quarter, vol. 19, no. 4, pp. 1–34, 2011.

[7] H. Zhang, Q. Du, M. Yao, and F. Ren, "Evaluation and clustering maps of groundwater wells in the red beds of chengdu, sichuan, china," Sustainability, vol. 8, no. 1, p. 87, 2016.

[8] N. Iqbal, F. Jamil, S. Ahmad, and D. Kim, "Toward effective planning and management using predictive analytics based on rental book data of academic libraries," IEEE Access, vol. 8, pp. 81 978–81 996, 2020.

[9] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). IEEE, 2013, pp. 1–7.

[10] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," Information sciences, vol. 275, pp. 314–347, 2014.

[11] T. Romary, F. Ors, J. Rivoirard, and J. Deraisme, "Unsupervised classification of multivariate geostatistical data: two algorithms," Computers & geosciences, vol. 85, pp. 96–103, 2015.

[12] J. E. Shortridge, S. D. Guikema, and B. F. Zaitchik, "Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds," Hydrology and Earth System Sciences, vol. 20, no. 7, pp. 2611–2628, 2016.

[13] S. Ahmad, F. Jamil, N. Iqbal, D. Kim et al., "Optimal route recommendation for waste carrier vehicles for efficient waste collection: A step forward towards sustainable cities," IEEE Access, vol. 8, pp. 77 875–77 887, 2020.

[14] S. Ahmad, N. Iqbal, F. Jamil, D. Kim et al., "Optimal policy-making for municipal waste management based on predictive model optimization," IEEE Access, vol. 8, pp. 218 458–218 469, 2020.

[15] S. Sahoo, T. Russo, J. Elliott, and I. Foster, "Machine learning algorithms for modeling groundwater level changes in agricultural regions of the us," Water Resources Research, vol. 53, no. 5, pp. 3878–3895, 2017.

[16] M. Gridach, "Character-level neural network for biomedical named entity recognition," Journal of biomedical informatics, vol. 70, pp. 85–91, 2017.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine, vol. 29, no. 6, pp. 82–97, 2012.

[18] F. Jamil, N. Iqbal, S. Ahmad, and D.-H. Kim, "Toward accurate position estimation using learning to prediction algorithm in indoor navigation," Sensors, vol. 20, no. 16, p. 4410, 2020.

[19] N. Iqbal, F. Jamil, S. Ahmad, and D. Kim, "A novel blockchain-based integrity and reliable veterinary clinic information management system using predictive analytics for provisioning of quality health services," IEEE Access, vol. 9, pp. 8069–8098, 2021.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," Applied energy, vol. 212, pp. 372–385, 2018.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] X. Song, Y. Liu, L. Xue, J. Wang, J. Zhang, J. Wang, L. Jiang, and Z. Cheng, "Time-series well performance prediction based on long short-term memory (lstm) neural network model," Journal of Petroleum Science and Engineering, vol. 186, p. 106682, 2020.

[24] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep lstm recurrent networks," Neurocomputing, vol. 323, pp. 203–213, 2019.

[25] D. Fan, H. Sun, J. Yao, K. Zhang, X. Yan, and Z. Sun, "Well production forecasting based on arima-lstm model considering manual operations," Energy, vol. 220, p. 119708, 2021.

[26] V. R. Rosa, E. Camponogara, and V. J. M. Ferreira Filho, "Design optimization of oilfield subsea infrastructures with manifold placement and pipeline layout," Computers & Chemical Engineering, vol. 108, pp. 163–178, 2018.

[27] H. Sahebi, S. Nickel, and J. Ashayeri, "Strategic and tactical mathematical programming models within the crude oil supply chain context—a review," Computers & chemical engineering, vol. 68, pp. 56–77, 2014.

[28] J. D. Kelleher, B. Mac Namee, and A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press, 2020.

[29] S. Emamgholizadeh, K. Moslemi, and G. Karami, "Prediction the groundwater level of bastam plain (iran) by artificial neural network (ann) and adaptive neuro-fuzzy inference system (anfis)," Water resources management, vol. 28, no. 15, pp. 5433–5446, 2014.

[30] C. Soares and K. Gray, "Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models," Journal of Petroleum Science and Engineering, vol. 172, pp. 934–959, 2019.

[31] P. C. Deka et al., "Support vector machine applications in the field of hydrology: a review," Applied soft computing, vol. 19, pp. 372–386, 2014.

[32] W. Sun and B. Trevor, "Combining k-nearest-neighbor models for annual peak breakup flow forecasting," Cold Regions Science and Technology, vol. 143, pp. 59–69, 2017.

[33] C. Hegde and K. Gray, "Use of machine learning and data analytics to increase drilling efficiency for nearby wells," Journal of Natural Gas Science and Engineering, vol. 40, pp. 327–335, 2017.

[34] Q. Tao, C. Gu, Z. Wang, and D. Jiang, "An intelligent clustering algorithm for high-dimensional multiview data in big data applications," Neurocomputing, vol. 393, pp. 234–244, 2020.

[35] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data mining for students' disposition analysis," Education and Information Technologies, vol. 23, no. 2, pp. 957–984, 2018.

[36] M.-H. Jeong, Y. Cai, C. J. Sullivan, and S. Wang, "Data depth based clustering analysis," in Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2016, pp. 1–10.

[37] N. Iqbal, S. Ahmad, D. H. Kim et al., "Towards mountain fire safety using fire spread predictive analytics and mountain fire containment in iot environment," Sustainability, vol. 13, no. 5, p. 2461, 2021.

[38] A. N. Khan, N. Iqbal, R. Ahmad, and D.-H. Kim, "Ensemble prediction approach based on learning to statistical model for efficient building energy consumption management," Symmetry, vol. 13, no. 3, p. 405, 2021.

[39] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, "Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6g-enabled maritime iot systems," IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5374–5385, 2020.

[40] F. Jamil, N. Iqbal, S. Ahmad, D. Kim et al., "Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid," IEEE Access, vol. 9, pp. 39 193–39 217, 2021.

[41] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan, "Adaptively constrained dynamic time warping for time series classification and clustering," Information Sciences, vol. 534, pp. 97–116, 2020.

[42] H. Li, J. Liu, K. Wu, Z. Yang, R. W. Liu, and N. Xiong, "Spatio-temporal vessel trajectory clustering based on data mapping and density," IEEE Access, vol. 6, pp. 58 939–58 954, 2018.

[43] S. Rani and G. Sikka, "Recent techniques of clustering of time series data: a survey," International Journal of Computer Applications, vol. 52, no. 15, 2012.

[44] M. Huang, Q. Bao, Y. Zhang, and W. Feng, "A hybrid algorithm for forecasting financial time series data based on dbscan and svr," Information, vol. 10, no. 3, p. 103, 2019.

[45] R. Garcia-Dias, S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Clustering analysis," in Machine Learning. Elsevier, 2020, pp. 227–247.

[46] M.-S. Yang and K. P. Sinaga, "A feature-reduction multi-view k-means clustering algorithm," IEEE Access, vol. 7, pp. 114 472–114 486, 2019.

[47] S. D. Kristjansson, A. Neudfeldt, S. W. Lai, J. Wang, D. Tremaine et al., "Use of historic data to improve drilling efficiency: a pattern recognition method and trial results," in IADC/SPE Drilling Conference and Exhibition. Society of Petroleum Engineers, 2016.

[48] M. He, N. Li, Z. Zhang, X. Yao, Y. Chen, and C. Zhu, "An empirical method for determining the mechanical properties of jointed rock mass using drilling energy," International journal of rock mechanics and mining sciences, vol. 116, pp. 64–74, 2019.

[49] C. Güler and G. D. Thyne, "Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c-means clustering," Water Resources Research, vol. 40, no. 12, 2004.

[50] T. Helstrup, N. O. Jørgensen, and B. Banoeng-Yakubo, "Investigation of hydrochemical characteristics of groundwater from the cretaceous-eocene limestone aquifer in southern ghana and southern togo using hierarchical cluster analysis," Hydrogeology Journal, vol. 15, no. 5, pp. 977–989, 2007.

[51] C. Güler, G. D. Thyne, J. E. McCray, and K. A. Turner, "Evaluation of graphical and multivariate statistical methods for classification of water chemistry data," Hydrogeology journal, vol. 10, no. 4, pp. 455–474, 2002.

[52] L. E. Widodo, T. A. Cahyadi, S. Notosiswoyo, and E. Widijanto, "Application of clustering system to analyze geological, geotechnical and hydrogeological data base according to hc-system approach," 2017.

[53] M. Fatehi and H. H. Asadi, "Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data, a case study from the dalli cu-au porphyry deposit in central iran," Ore Geology Reviews, vol. 81, pp. 245–255, 2017.

[54] C. Reimann, P. Filzmoser, R. Garrett, and R. Dutter, Statistical data analysis explained: applied environmental statistics with R. John Wiley & Sons, 2011.

[55] J. M. Morrison, M. B. Goldhaber, K. J. Ellefsen, and C. T. Mills, "Cluster analysis of a regional-scale soil geochemical dataset in northern california," Applied Geochemistry, vol. 26, pp. S105–S107, 2011.
[56] Z. Deng, Y. Hu, M. Zhu, X. Huang, and B. Du, "A scalable and fast optics for clustering trajectory big data," Cluster Computing, vol. 18, no. 2, pp. 549–562, 2015.
[57] 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, 2017.
[58] Y. Zhao, A. Noorbakhsh, M. Koopialipoor, A. Azizi, and M. Tahir, "A new methodology for optimization and prediction of rate of penetration during drilling operations," Engineering with Computers, vol. 36, no. 2, pp. 587–595, 2020.
[59] M. Bataee, S. Irawan, and M. Kamyab, "Artificial neural network model for prediction of drilling rate of penetration and optimization of parameters," Journal of the Japan Petroleum Institute, vol. 57, no. 2, pp. 65–70, 2014.
[60] C. Hegde and K. Gray, "Evaluation of coupled machine learning models for drilling optimization," Journal of Natural Gas Science and Engineering, vol. 56, pp. 397–407, 2018.
[61] C. Hegde, H. Daigle, K. E. Gray et al., "Performance comparison of algorithms for real-time rate-of-penetration optimization in drilling using data-driven models," SPE Journal, vol. 23, no. 05, pp. 1–706, 2018.
[62] O. S. Ahmed, A. A. Adeniran, and A. Samsuri, "Computational intelligence based prediction of drilling rate of penetration: A comparative study," Journal of Petroleum Science and Engineering, vol. 172, pp. 1–12, 2019.
[63] S. B. Ashrafi, M. Anemangely, M. Sabah, and M. J. Ameri, "Application of hybrid artificial neural networks for predicting rate of penetration (rop): A case study from marun oil field," Journal of Petroleum Science and Engineering, vol. 175, pp. 604–623, 2019.
[64] 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 Science Informatics, vol. 12, no. 3, pp. 319–339, 2019.
[65] S. Tewari and U. Dwivedi, "Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs," Computers & Industrial Engineering, vol. 128, pp. 937–947, 2019.
[66] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102–4107, 2017.
[67] C. Thirumalai, R. Kanimozhi, and B. Vaishnavi, "Data analysis using box plot on electricity consumption," in 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 2. IEEE, 2017, pp. 598–600.
[68] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," Epidemiology (Cambridge, Mass.), vol. 21, no. 1, p. 128, 2010.

NAEEM IQBAL is currently pursuing Ph.D. in the Department of Computer Engineering at Jeju National University, the Republic of Korea. He received his MS in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan in 2019. He did his BS in Computer Science from the COMSATS University Islamabad, Attock Campus. He has professional experience in the software development industry and in academic as well. His research work mainly focused on Machine Learning, Big Data, AI-based Intelligent Systems, Analysis of Optimization Algorithms, and Blockchain-based Applications.

ATIF RIZWAN is currently pursuing Ph.D. in the Department of Computer Engineering at Jeju National University, the Republic of Korea. He received his MS in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan in 2020 and he has also completed his Master of Computer science (16 years) from the COMSATS University Islamabad, Attock Campus. He has good industry experience in software development and testing. His research work focused on machine learning, Data and Web Mining, analysis of optimization of core algorithms and IoT based applications.

ANAM NAWAZ KHAN is currently pursuing Ph.D. in the Department of Computer Engineering at Jeju National University, the Republic of Korea. She received his MS in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan in 2019. She did his BS in Computer Science from the COMSATS University Islamabad, Attock Campus. Her research work mainly focused on Machine Learning, Data Mining, and Energy Prediction Systems, and IoT 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. His research work is focused on the application of prediction and optimization algorithms to build IoT-based solutions. His research interests mainly focused on Machine Learning, Data Mining, related applications.

BONG WAN KIM is working as a researcher in the Urban Space ICT Lab, the Republic of Korea. His research work mainly focused on, Middleware technology, device control object Internet technologies, intelligent object technology, equipment identity management technologies, Low - power wireless communication technology, low-power sensor control technology, and low-power device operation technology.

KWANGSOO KIM is working as a researcher in the Urban Space ICT Lab, the Republic of Korea. His research work mainly focused on, Sensing data collection, data analysis, data management and open interfaces, data-based situational awareness, Neural network modeling, neural network learning, the neural network inference and artificial intelligence algorithms, Space modeling data, spatial analysis, data visualization, data processing algorithm space, space Big Data Management.

DOHYEUN KIM received the B.S. degree in electronics engineering from Kyungpook National University, South Korea, in 1988, and the M.S.and Ph.D. degrees in information telecommunication from Kyungpook National University, South Korea, in 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, South 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 research interests include sensor networks, M2M/IOT, energy optimization and prediction, intelligent service, and mobile computing.