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

Groundwater Level Prediction Model using Correlation and Difference Mechanisms based on Boreholes Data for Sustainable Hydraulic Resource Management

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

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

Corresponding author: Do-Hyeun 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 Drilling data for groundwater extraction incur changes over time due to variations in hydrogeological and weather conditions. At any time, if there is a need to deploy a change in drilling operations, drilling companies keep monitoring the time-series drilling data to make sure it is not introducing any changes or new errors. Therefore, a solution is needed to predict groundwater levels (GWL) and detect a change in boreholes data to improve drilling efficiency. The proposed study presents an ensemble GWL prediction (E-GWLP) model using boosting and bagging models based on stacking techniques to predict GWL for enhancing hydraulic resource management and planning. The proposed research study consists of two modules; descriptive analysis of boreholes data and GWL prediction model using ensemble model based on stacking. First, descriptive analysis techniques, such as correlation analysis and difference mechanisms, are applied to investigate boreholes log data for extracting underlying characteristics, which is critical for enhancing hydraulic resource management. Second, an ensemble prediction model is developed based on multiple hydrological patterns using robust machine learning (ML) techniques to predict GWL for enhancing drilling efficiency and water resource management. The architecture of the proposed ensemble model involves three boosting algorithms as base models (level-0) and a bagging algorithm as a meta-model that combines the base models predictions (level-1). The base models consist of the following boosting algorithms; eXtreme Gradient Boosting (XGBoost), AdaBoost, Gradient Boosting (GB). The meta-model includes Random Forest (RF) as a bagging algorithm referred to as a level-1 model. Furthermore, different evaluation metrics are used, including mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE), mean absolute percentage error (MAPE), and R2 score. The performance of the proposed E-GWLP model is compared with existing ensemble and baseline models. The experimental results reveal that the proposed model performed accurately in respect of MAE, MSE, and RMSE of 0.340, 0.564, and 0.751, respectively. The MAPE and R2 score of our proposed approach is 12.658 and 0.976, respectively, which signifies the importance of our work. Moreover, experimental results suggest that E-GWLP model is suitable for sustainable water resource management and improves reservoir engineering.

INDEX TERMS Groundwater Level Prediction, Machine Learning, Bagging and Boosting, Correlation Analysis, Time-Series Data

I. INTRODUCTION

Due to variation in climatic conditions, overexploitation of groundwater, and lack of sustainable management of groundwater resources results in a repid increase in water supply crises [1], [2]. Groundwater is a scarce unseen water resource in natural reservoirs in soil or rocks beneath the earth's surface [3]. Groundwater plays a vital role in fulfilling the requirements of industrial development, economic growth of the country, and providing safe water to living beings worldwide [4], [5]. However, in recent years, it is decreasing due to improper groundwater resource extraction and overexploitation [6]. Drilling is widely considered to extract groundwater resources to fulfill the needs of living beings. Increased groundwater demand and its exploitation have surged the drilling process for groundwater extraction. Drilling and extraction of groundwater may lead to a decline in groundwater resources, increased boreholes depth, and higher drilling costs [7]. The drilling process for groundwater level has some significant risks and complexities concerning economy, environment, and sustainability [8].

Drilling boreholes to gain the GWL is a complicated process that accounts for a massive amount of budgets due to dynamic variations in hydrogeological characteristics. Factors influencing the cost of the drilling process involve the type of soil, land layer, boreholes depth, intended use, machinery, skilled workforce, and materials needed [9]. Hence, drilling depth prediction is crucial for improvements in the overall drilling process, holistic management of hydraulic resources, development of city, underground safety, risk assessment, etc. However, GWL prediction is a complex and dynamic process due to variations in hydrogeological properties. Unfortunately, none of the existing work has achieved reliable prediction accuracy due to complex parameters influencing boreholes depths [10]. Proper utilization of time-series analysis of boreholes log data and mathematical tools can help to predict GWL for enhancing the efficiency of future boreholes.

Groundwater plays a very vital role in the irrigation and food production of a country [11]. Groundwater usage has grown enormously during the past few decades. One of the primary reasons is the advancement in drilling technologies [12]. Increased water usage has surged the demand for drilling groundwater. Due to rapid climatic and geological changes, the prediction of groundwater-related aspects has become difficult. Therefore time-series analysis of ground-level data and future trend prediction for land subsidence is immensely beneficial for achieving sustainability and efficient use of resources. Time series analysis of groundwater level data will aid the detection of trends and patterns, and behaviors for the identification of declining water levels. Time series modeling provides a better fitting model as compared to other groundwater level data models.

Noisy and varying time-series boreholes data has made it a challenging process to search and locate differences and dissimilarities in time-series data in a large context. There is a lack of such efficient systems and techniques to handle the huge amount of available data to improve the drilling process [13]. Time-series boreholes data possess a high dimensionality resulting in slower access times and high computational complexity. The keynote is the fast search of real time-series data set and the difficulty with time-series because we cannot precisely apply string match and directly index time-series. Therefore employ distance functions and a much fast algorithm than a simple linear scan. Furthermore, it becomes computationally expensive in terms of cost (time and storage) to apply analysis techniques to the original borehole's time-series data. Difference functions are undoubtedly significant for time-series modeling and prediction. Because it is not practical to apply machine learning techniques on raw and un-preprocessed time series data. Therefore, it is needed a higher-level representation of data for efficient computation and extraction of higher-order features. A vast amount of methods exist for generating a difference between time-series data; these methods include Discrete Fourier Transform (DFT) [14], Discrete Wavelet Transform (DWT) [15], piecewise aggregate approximation [16], 1-lag difference algorithm [17], to name a few.

Drilling process of boreholes generates a vast amount of boreholes log data. There are various sources to acquire boreholes data, starting from drilling activity breakdown, soil colors, land layers, geology and casing information, bottom hole Assembly, and bit information [18]. An essential feature of the borehole's time-series data is high dimensionality and dynamicity. The speed at which the boreholes data is growing does not match the corresponding development techniques of data interpretation and analysis [19]. At present, the drilling industry faces a major challenge in finding ways to tackle such huge volumes of boreholes data for analysis and modeling. The ability to measure the differences between instances is crucial to various data mining applications. We can define time series as composed of complex data objects found in many applications like the stock market, hydro-geology, medicine, telecommunication, etc. The enormous increase in data generating and collecting devices has resulted in the construction of time-series databases. Time-series data analysis and evaluation techniques are highly demanded by data scientists for comparing values, trends, patterns, and periodicity.

With the development of robust time-series models, it is quite possible to develop efficient ML models using time-series boreholes data. In recent years technological advancements in ML have brought breakthrough changes concerning efficient data processing and data mining solutions such as XGBoost, Artificial Neural Networks (ANN), Deep Learning (DNN), and Support Vector Regression (SVR). All these powerful techniques have facilitated improvement in the prediction performance of complex time-series data. ML techniques have been widely utilized in many areas, such as regression [20], classification [21], [22], patterns mining [23], [24], decision-making systems [25], [26], to name of a few. ML-based approaches tend to produce more robust predictions than conventional methods due to their ability to limit uncertainty concerning input variables having various nonlinear dependencies to generate accurate and reliable predictions. Therefore, in this research study, we employed ML-based ensemble and conventional techniques to predict GWL for sustainable water resource management.

The notable contributions of our proposed work are:

- The notable contribution of our proposed work is to employ data and predictive analytics to predict GWL for sustainable water resource management.
- Integrating boosting and bagging models using stacking technique to develop an ensemble prediction model to predict GWL for facilitating hydraulic management for sustainable groundwater resources.
- Different descriptive analyses are utilized to investigate boreholes time-series data for extracting underlying hydrogeological patterns. The descriptive analysis includes boreholes log data analysis based on borehole depth, analysis of boreholes data according to soil color patterns, rock unit, stratum layer, to name a few.
- Different hydrogeological parameters are computed from the historical boreholes log data; total borehole's depth, total number of days spent on each borehole, core soil color, core rock layer, and core stratum layer.
- Detailed comparative study is illustrated to signify the significance of the E-GWLP model compared to the existing baseline models.

The rest of the paper is summarized as follow. Section II presents a detailed review of the existing GWL prediction models; Section III describes proposed methodology of the E-GWLP model. Section IV describes boreholes log data. Section V presents data preprocessing, descriptive data analysis, and features extraction modules. Section VI presents proposed difference mechanism to detect change in time series data. In section VII, implementation and experimental environment are discussed. Section VIII presents prediction results and analysis. Section IX presents conclusion of the proposed E-GWLP model.

II. LITERATURE REVIEW

In this section, a detailed survey is conducted to highlight the strengths and weaknesses of the existing GWL prediction models. GWL prediction is considered one of the challenging tasks due to improper extraction, dynamic variations in hydrogeological properties, and over-exploitation [6]. Recently, different ML and mathematical models are suggested by different researchers to predict GWL [27], [28]. Existing prediction models have been developed to match the complexities and accuracy of estimation of GWL due to different hydrogeological and structural properties [27], [29]. In the last few years, most of the research studies used soft computing techniques for GWL [27]. These soft-computing techniques included ANN [30], support vector machines (SVM) [31], and adaptive neuro-fuzzy interface systems (ANFIS) [32].

The aforementioned soft-computing techniques have been widely used to predict hydrological parameters due to multiple factors, such as low computational complexity, high precision, fast training, fast performance time, to name a few [33]. For instance, in [34], the authors developed a hybrid prediction model based on ANN and wavelet theorem to predict GWL in Canada. The authors modeled fluctuations in GWL based on monthly recorded temperature. In [35], the authors developed and compared feed-forward ANN with the conventional regression model for estimating GWL in the time interval of 1 hour. In [36], ANN and ANFIS models are developed to simulate and predict GWL in Iran. The authors considered the following three parameters as an input; a flow of irrigation returned, prediction rate, and pumping rate, to train and test ANN and ANFIS models. The results revealed that the ANFIS model performed accurately compared to the ANN. Another study presented in [37] applied ANN and SVM techniques to predict GWL prediction based on boreholes data acquired from 5 stations in Republic of Korea. The results indicated that the SVM model was more precise and accurate compared to the conventional ANN model. Furthermore, a study presented in [38] utilized ANN and SVM to predict water table depth.

In the last few years, other ensemble and conventional ML models are also developed to predict GWL prediction for sustainable water resource management [39], [40]. In [39] the authors presented an ensemble model based on KNN and RF for three months ahead of groundwater table prediction based on seasonal changes. In [40], the authors proposed an enhanced RF prediction model based on the combination of random features to forecast GWL using two features; temperature (Celsius) and precipitation (Millimeters). The authors reported that the R2 score value of the enhanced RF is 0.8223 for long-term forecasting, which is still improvable. RF model can be efficiently used to handle small and large datasets [41]. It is a robust ML model that produces better generalization to overcome overfitting issues for modeling applications related to hydrology [42]. The authors developed an enhanced RF model to forecast GWL in data-scarce regions [40]. A detailed comparative study is presented in [43] to explain a wide range of RF applications in the field of hydrogeology. Another study presented in [44] also implemented RF using a geographic information system (GIS) based on potential mapping for predicting groundwater level. The authors developed potential maps that can be applied to underground resource exploration. In [45], the authors developed a classification model based on RF to predict the layer to extract underground water samples. The classification model was developed based on the main ion composition of the underground water samples. Efficient modeling of boreholes log data is vital for sustainable hydraulic resource development and management. In [46], a prediction model was developed based on RF mode to predict water level variations of the lake for sustainable development. The experimental results of the RF model were compared with existing ML models; ANN, SVM, and linear regression (LR) models.

Likewise, statistical techniques are employed to predict

GWL based on time-series data. These methods have been proposed for evaluating temporal trends concerning groundwater like regression analysis to complex parametric and non-parametric techniques. One of the drawbacks of using a simple regression model is its inability to handle non-linear patterns [47]. Frequently used time-series prediction models include autoregressive integrated moving average (ARIMA), regression analysis and exponential smoothing. In [47], the authors proposed a non-parametric approach (Mann-Kendall) for analysis of trends in groundwater level. Another study used a geostatistical approach to predict spatial and temporal groundwater variation using ARIMA and sequential Gaussian simulation method [48]. In [49], the authors employed time series modeling to forecast fluctuations in groundwater levels. Likewise, predicted groundwater levels using integrated time series, ARIMA, and Holt-Winters exponential smoothing (HWES). However, experimental results show superior performance by the HWES approach. For trend analysis, a new approach called innovative trend analysis (ITA) based on a statistical method is used by many researchers. ITA performs a comparative analysis of time series data without considering statistical assumptions [50]. Furthermore, in [51], a novel method was proposed for identification of trends their magnitude for groundwater levels involving temperate climatic conditions for efficient management of scarce water resources.

Time-series is a sequence of random variables across time stamps upon which we apply tools and mathematical models to achieve the desired goal. Time-series analysis has been frequently reported in the literature for prediction with varying complexities and accuracies [52], [53]. Prediction of time-series involves predicting future data points based on historical data such that the error is minimized. Finding differences between time-series datasets is an integral component of the development process. Comparison of data enables us to locate differences and make our analysis more comprehensive. Moreover, we can check the variables that caused the difference [54]. The basic goal of difference algorithms is to deliver an efficient strategy for generating differences. Due to external events, the time-series borehole's data is subjected to interruption. A difference is created in pre-and post-intervention stages, which may be temporary or permanent.

However, a plethora of prediction and difference mechanisms are available in the literature to predict and compare time series. Differentiating data can also be done using various test types like parametric and nonparametric, for example, a distribution-free test where no information about the distribution of the population is given lie under the category of parametric test it uses qualitative data, e.g., Wilcoxon, Mann Whitney, and Kruskal-Wallis tests. In the parametric test case, a normal distribution is considered, e.g., t-test and ANOVA [9]. There are several difference/dissimilarity measures employed in various studies for comparison of time-series data. Following are some statistical methods for finding differences between time series data. This includes T-test [55] that deals with parametric data and makes a comparison between two-time series, the virtual classifier (VI) [56] for interpreting change that occurs in two consecutive windows, Rank Preservation [57] for comparing two matrices by taking column-wise correlation, CUSUM also called as cumulative sum test [58] for detection of change points in a time series, Spearman correlation [59] for measuring association among two data groups, ANOVA test [60] make a comparison of more than three paired data groupings,

To the best of our knowledge, many existing prediction models were developed based on the conventional ANN algorithm to predict GWL. Some of the existing models were implemented based on SVM and ANN to forecast GWL. However, still, these models did not achieve accurate prediction results due to variations in hydrogeological patterns. This study aims to develop an ensemble model by integrating boosting and bagging models using a stacking combinator to predict GWL sustainable hydraulic planning and management. Furthermore, descriptive data analysis techniques are utilized to analyze the hydrogeological patterns of time-series boreholes data acquired from Jeju National University (JNU), Republic of Korea. Moreover, different hydrological and time-series patterns are extracted from real boreholes data to evaluate and compare the proposed model with baseline ensemble and ML models. Therefore, to the best of the author's knowledge, it is the first attempt to integrate boosting and bagging models to develop a robust E-GWLP model based on hidden hydrological characteristics for sustainable water management.

III. METHODS

This section presents a detailed methodology of the proposed E-GWLP model. The proposed E-GWLP model aims to utilize sophisticated and robust ML ensemble approaches to improve hydraulic resource management.

A. PROPOSED MODEL OVERVIEW

An overview of the proposed E-GWLP model is described. Fig. 1 exhibits the block diagram to analyze the detailed overview of our proposed method. The block diagram describes the functional flow of the proposed model. The functional flow of the proposed model consists of various steps. In step 1, raw data of the boreholes-log is passed to the preprocessing module. Step 2 indicates preprocessing module that aims to preprocess raw data by removing irrelevant features, handling missing values, and label encoding to increase the efficiency of the boreholes-log data. Next, in step 3, preprocessed data is passed to the features engineering module to construct new features using the existing preprocessed features set. Data analysis is considered an integral module in data mining to investigate the underlying characteristics of the historical data. Therefore, in step 4, the data analysis module is presented to perform different types of analysis, including time-series analysis, statistical analysis, etc. Features selection is an important process to reduce a large feature space by eliminating the least contributed features without

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.2021.3094735, IEEE Access

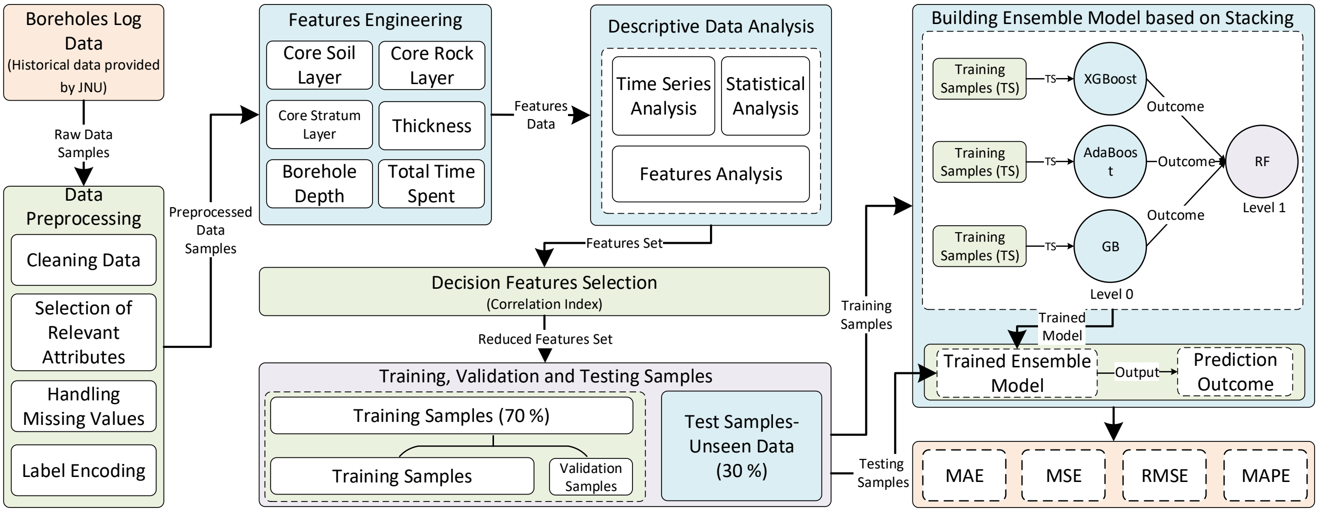


FIGURE 1. Block diagram of the proposed E-GWLP model.

losing the accuracy and efficiency of the proposed model. Step 5 presents the features selection process to select the most promising features from the base and derived features. In step 6, the data splitting module divides reduced feature data into training and testing samples. In step 7, an ensemble prediction model is trained using training samples based on the stacking technique. Similarly, testing samples are utilized to evaluate the trained ensemble prediction model to determine the efficiency of the proposed model. Finally, in step 8, different error estimation measures are considered to check the prediction error of the developed model.

B. PROPOSED ARCHITECTURE OF E-GWLP MODEL

This subsection depicts the main architecture for developing E-GWLP model. Fig. 2 introduces the layered architecture of the proposed model for predicting GWL to improve the development of hydraulic resource management for future groundwater extraction. This layered architecture of the proposed model architecture consists of 5 layers. The first layer presents time-series boreholes data acquired from JNU, Republic of Korea. The boreholes log dataset consists of the following attributes, including borehole ID, altitude, soil color patterns, rock units, strata codes, etc. The second layer presents data preprocessing and analysis of the boreholes data. The acquired boreholes log data is not in reliable format; therefore, cleaning of raw data is required to convert unprocessed data into a meaningful form for data mining (DM). Therefore, data processing is taken into account to remove irrelevant attributes and other outliers from the acquired data. Data analysis takes preprocessing data as an input to process and investigate trends of the historical boreholes log data. Different hydrogeological and time interval analyses are conducted to analyze underlying characteristics of the preprocessed boreholes log data, which can be considered helpful for the future drilling process. In the third layer, difference mechanisms are developed based on lag-1 difference and unsupervised difference algorithms to detect seasonality change in time-series data observations. The fourth layer presents a proposed ensemble prediction model using level 0 and level 1 models based on stacking to predict GWL. One of the primary objectives of our work is to integrate boosting and bagging models using stacking to build an ensemble model for predicting GWL. Furthermore, different conventional ensemble and baseline ML models are also developed. Lastly, different prediction error metrics are implemented to measure the prediction error of the E-GWLP model. The prediction error of the proposed E-GWLP is also compared with state-of-art and traditional hybrid models to signify the importance of the proposed work.

C. FLOW DIAGRAM OF THE PROPOSED E-GWLP MODEL

In this subsection, a detailed flow of our proposed E-GWLP model is exhibited in Fig.3. The functional flow of our proposed method consists of the following steps; collection of boreholes log data, preprocessing of collected data, descriptive analysis of boreholes log data, extraction of hydrogeological features, normalization of decision features, utilization of difference mechanisms, developing ensemble model, and performance evaluation. The boreholes data contains 9,287 data samples for boreholes of different regions in the Republic of Korea.

The dataset includes 12 input features; borehole log ID, altitude, geographic coordinates X and Y, starting (top) depth, ending (bottom) depth, thickness, standard Korean layer name, starting and ending drilling date, and groundwater level. The acquired dataset contains irrelevant data and outliers; therefore, data preprocessing techniques are used to clean and filter out trivial features to accumulate the consistency of the dataset. Next, preprocessed data are

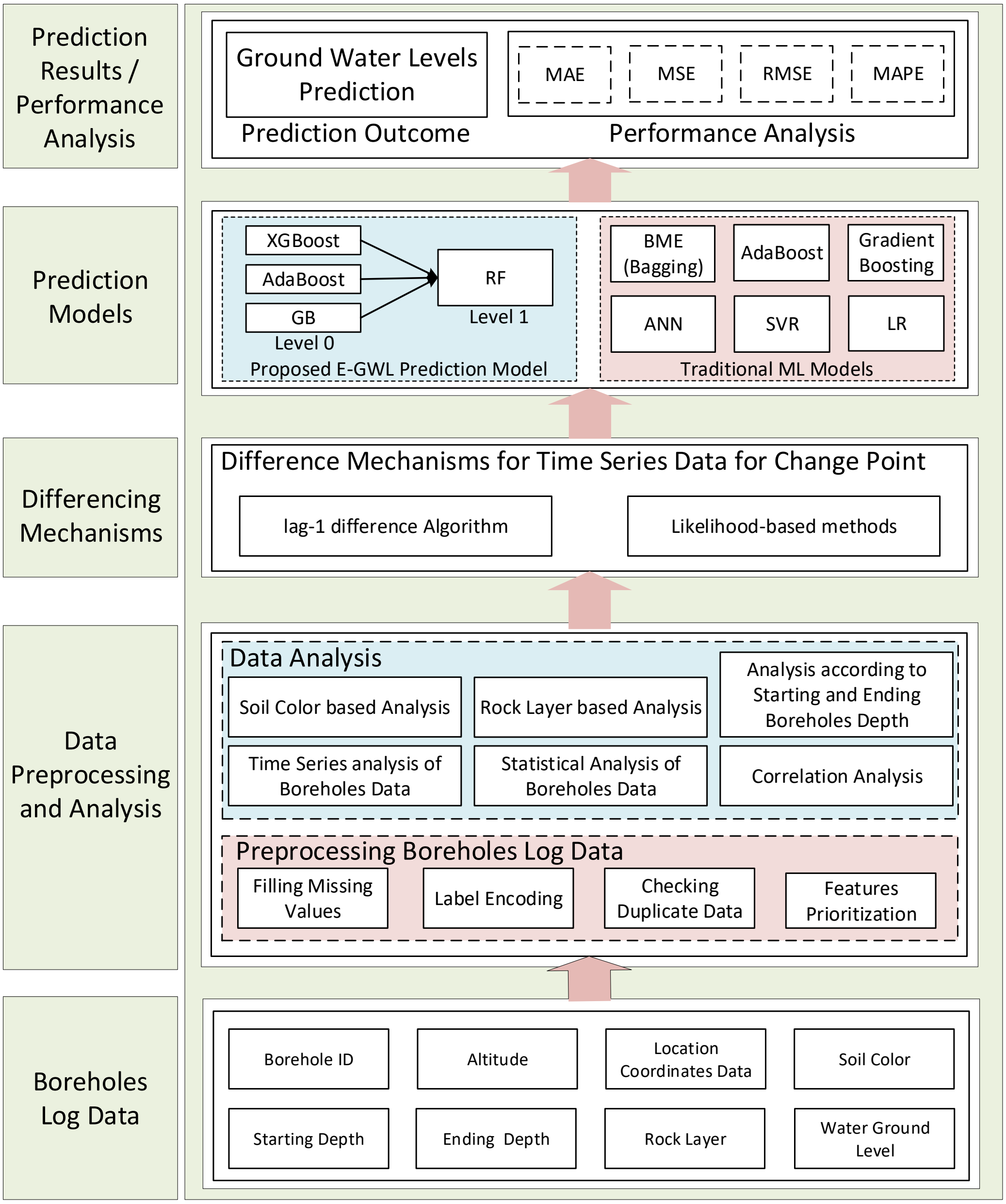

Prediction Results / Performance Analysis
Ground Water Levels Prediction
Prediction Outcome
MAE
MSE
RMSE
MAPE
Performance Analysis
Prediction Models
XGBoost
AdaBoost
GB
Level 0
RF
Level 1
Proposed E-GWL Prediction Model
BME (Bagging)
AdaBoost
Gradient Boosting
ANN
SVR
LR
Traditional ML Models
Differencing Mechanisms
Difference Mechanisms for Time Series Data for Change Point
lag-1 difference Algorithm
Likelihood-based methods
Data Preprocessing and Analysis
Data Analysis
Soil Color based Analysis
Rock Layer based Analysis
Analysis according to Starting and Ending Boreholes Depth
Time Series analysis of Boreholes Data
Statistical Analysis of Boreholes Data
Correlation Analysis
Preprocessing Boreholes Log Data
Filling Missing Values
Label Encoding
Checking Duplicate Data
Features Prioritization
Boreholes Log Data
Borehole ID
Altitude
Location Coordinates Data
Soil Color
Starting Depth
Ending Depth
Rock Layer
Water Ground Level

FIGURE 2. Proposed architecture of the E-GWLP model.

passed to the data analysis and features extraction modules. The preprocessed data are analyzed based on different data analysis techniques to highlight the trends and patterns of the historical time-series boreholes data. Furthermore, different hydrogeological and time interval features are computed from the preprocessed boreholes log data; days spent on each borehole drilling, total drilling depth, soil color with maximum borehole depth, to name a few. Next, data normalization technique is implemented to scale down feature values in uniform range [0,1]. Correlation and difference mechanisms are applied to evaluate the linear relationships of the decision variables and identify a change in time-series observations. In the next step, an ensemble model is developed based on the combination of boosting and bagging models using stacking to predict GWL. The proposed ensemble model is formed based on the integration of two models; base and meta models. The base models are developed based on three boosting models: XGBoost, AdaBoost, and GB. Similarly, a meta-model involves an RF model as a bagging algorithm to learn from the base model predictions. The prediction outputs of the base models are fused to the meta-model as input to learning from these predictions. The stacking method is used

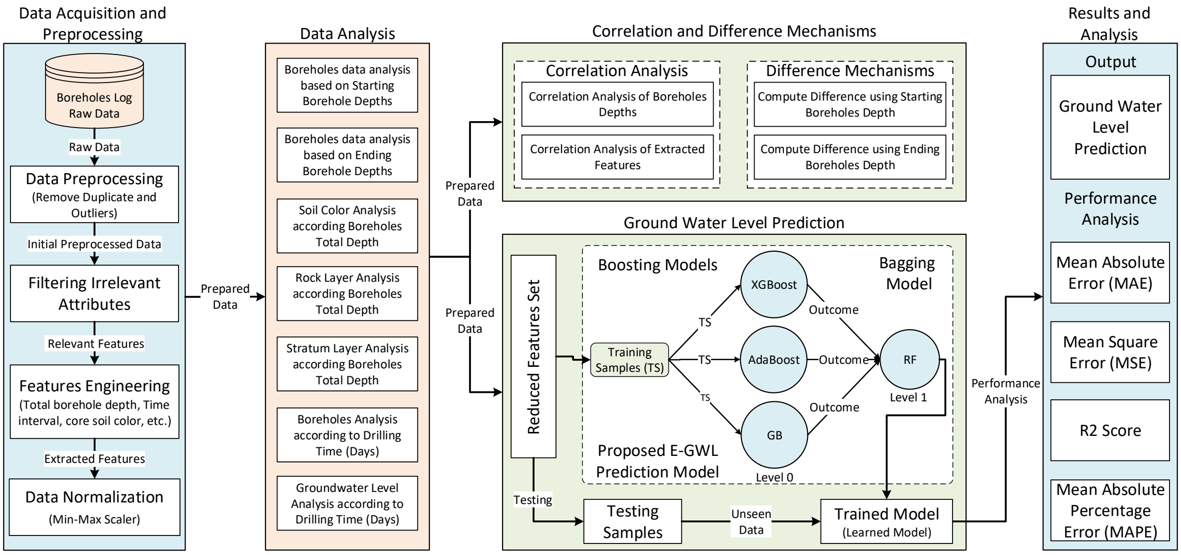

Data Acquisition and Preprocessing
Boreholes Log Raw Data
Raw Data
Data Preprocessing (Remove Duplicate and Outliers)
Initial Preprocessed Data
Filtering Irrelevant Attributes
Relevant Features
Features Engineering (Total borehole depth, Time interval, core soil color, etc.)
Extracted Features
Data Normalization (Min-Max Scaler)
Prepared Data
Data Analysis
Boreholes data analysis based on Starting Borehole Depths
Boreholes data analysis based on Ending Borehole Depths
Soil Color Analysis according Boreholes Total Depth
Rock Layer Analysis according Boreholes Total Depth
Stratum Layer Analysis according Boreholes Total Depth
Boreholes Analysis according to Drilling Time (Days)
Groundwater Level Analysis according to Drilling Time (Days)
Prepared Data
Prepared Data
Correlation and Difference Mechanisms
Correlation Analysis
Correlation Analysis of Boreholes Depths
Correlation Analysis of Extracted Features
Difference Mechanisms
Compute Difference using Starting Boreholes Depth
Compute Difference using Ending Boreholes Depth
Ground Water Level Prediction
Reduced Features Set
Boosting Models
Bagging Model
Training Samples (TS)
TS
TS
TS
XGBoost
AdaBoost
GB
Outcome
Outcome
Outcome
RF
Level 1
Level 0
Proposed E-GWL Prediction Model
Testing
Testing Samples
Unseen Data
Trained Model (Learned Model)
Performance Analysis
Results and Analysis
Output
Ground Water Level Prediction
Performance Analysis
Mean Absolute Error (MAE)
Mean Square Error (MSE)
R2 Score
Mean Absolute Percentage Error (MAPE)

FIGURE 3. Detailed flow model of the proposed ensemble prediction model.

as a combinator to combine base and meta models to draw a conclusion. The prediction results of the proposed ensemble model are evaluated using different evaluation measures; MAE, MSE, RMSE , MAPE, to name a few. Furthermore, prediction results of our E-GWLP model are compared with baseline approaches to signify the usefulness and relevance of the proposed research study.

IV. TIME-SERIES BOREHOLES DATA PRESENTATION

This section presents a boreholes log data provided by the JNU, Republic of Korea. The considered boreholes log data consists of 9,287 samples along with 1,987 unique boreholes. The collected data includes following data features, such as borehole log ID, geographic coordinates, starting depth of thickness layer, ending depth of thickness layer, rock unit, patterns of soil color, groundwater level, etc. Groundwater level represents the depth under the earth's surface that is permeated with water. Soil color represents the color patterns of soil under the ground. Stratum layer is defined as a layer of sedimentary rock that formed under the ground surface. The land layer represents the rock unit under the ground surface; it can be classified as igneous or sedimentary rocks. The detailed summary of the boreholes log data is presented in Table 1.

TABLE 1. Summary of the boreholes log data.

#	Data Attribute	Description
1	Borehole Log ID	An identifier to represents borehole uniquely.
2	Resonance of Borehole	It represents resonance ID of borehole in the given geographic region.
3	Location Coordinates	Location coordinates represents borehole location in the selected region.
4	Altitude	It indicates altitude of the drilling process.
5	Academic Strata Layer	It indicates rock unit under the ground surface.
6	Soil Color	It represents color of the soil below the earth surface
7	Land Layer	It includes rock layer and other layers at the earth surface.
8	Starting Borehole-Log Depth	Starting depth of thickness layer for borehole k at time t.
9	Ending Borehole-Log Depth	Ending depth of thickness layer for borehole k at time t.
10	Starting Borehole-Log Date	It is starting date at which borehole-log process begin.
11	Ending Borehole-Log Date	It is ending date at which borehole-log process begin.
12	Ground Water Level	It is represents depth of the earth surface in which rock and soil layers are saturated with water.

V. DATA PREPROCESSING, DESCRIPTIVE ANALYSIS AND FEATURES EXTRACTION

This section describes collection of boreholes log data, cleaning of boreholes log data, and descriptive analysis to investigate underlying characteristics of drilling process.

A. PREPROCESSING OF DRILLING DEPTH DATA

Data preprocessing is a vital and challenging task in DM to clean and prepare preprocessed data model. Data preprocessing model aims to reduce the dataset size, determine the relation between data attributes, normalize data to get uniformity, remove noise and outliers, to name a few. It also helps to increase the consistency of the dataset, reduce computational and storage costs. However, unclean data will significantly affect data-driven methods and led to poor results. Therefore, it is required to clean raw data to find outliers and missing value attributes. In this study, several steps are carried out to convert raw data into a reliable format.

1) The given drilling dataset is processed to find records having missing value attributes. To fill missing values of attributes, a central tendency method is used to fill missing values to enhance reliability.
2) Duplicate boreholes log samples are highlighted and removed from the time-series boreholes dataset to enhance the consistency of the dataset.
3) All those borehole records are highlighted from the boreholes dataset, which doesn't have soil color and land layer values.
4) Static and irrelevant features are removed from the dataset to reduce storage cost and computational complexity.
5) All other data outliers are removed from the dataset that causes inconsistency issues.
6) Ordinal encoding method is used to transform categorical variables into continuous variables by assigning a unique integer to each category of categorical variables.

B. DESCRIPTIVE DATA ANALYSIS

Data analysis is a systematic process of applying statistical and logical methods to unearth hidden characteristics of the prepared dataset. Data analysis aims to discover hidden patterns and useful information from a massive amount of data to draw conclusions. Therefore, a preprocessed drilling depth data is used to apply descriptive data analysis techniques to track historical data for underground water characteristics. Different descriptive analyses are performed to track and discover hidden patterns and characteristics from the drilling depth data, which is essential for sustainable water resource management.

Fig. 4 depicts drilling data based on starting drilling depth frequency. Along the y-axis, we have starting depth frequency, and on the x-axis, borehole code is plotted. It can be observed that starting drilling depth frequency data fluctuated between the limits of 0 to 70 meter. The minimum starting drilling depth is 0 meter; whereas maximum starting drilling depth for drilling location is 70 meter. The starting drilling depth varies between a minimum and maximum drilling depth for the given drilling locations.

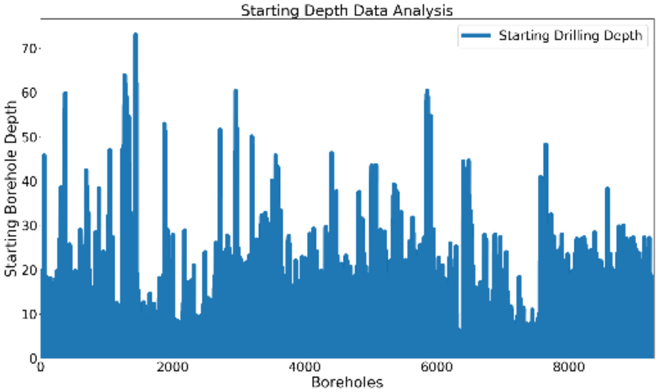


FIGURE 4. Descriptive analysis of the borehole based on starting drilling depths.

Similarly, Fig. 5 examines boreholes log data based on ending (bottom) drilling depth frequency. It can be observed that ending drilling depth frequency data fluctuated between values 0 to 75 meter for drilling locations at time t. The x-axis represents drilling locations, and the y-axis represents drilling depth for location x at time t. A major rise in ending depth frequency can be seen for boreholes between 0 and 2,000, while the rest of borehole codes show fewer fluctuations comparatively between borehole codes 4,000 and 8,000 and above.

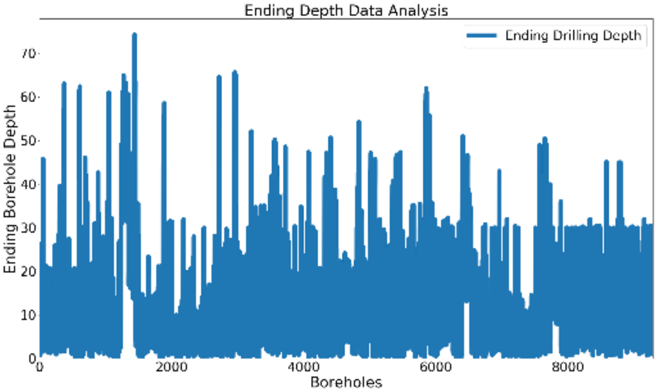


FIGURE 5. Descriptive analysis of the borehole based on ending drilling depths.

Fig. 6 presents a comparative analysis in order to compare the starting and ending drilling depth frequency. Along y-axis drilling depth frequency is plotted against borehole codes on the x-axis. It can be observed that starting drilling depth frequency data fluctuated between values 0 to 70 meter. Whereas, it can be observed that ending drilling depth frequency data fluctuated between 0 and 75 meter. The decline in groundwater affects the drilling depth frequency, which is evident from the starting and ending drilling depth.

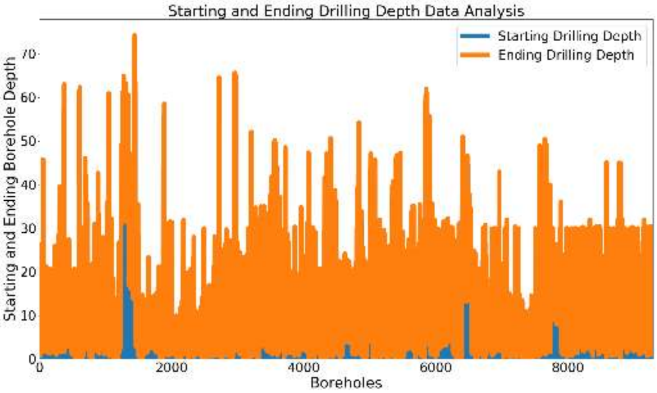


FIGURE 6. Comparison of the borehole based on starting and ending drilling depths.

In Fig. 7, average and maximum boreholes depth is analyzed for each unique pattern of soil color. The analysis investigates that the average and maximum boreholes depth varies due to the structure of the rock types. The analysis shows that the maximum average boreholes depth is 20.58 meters for soil color “Tan” among listed soil colors. Furthermore, soil color "Light Brown" has minimum average boreholes depth of 15.32 meters. Similarly, maximum analysis depicts that soil color pattern "Partridge" has maximum borehole depth of 74.28 meters, which indicates that the drilling process is difficult compared to the other soil colors.

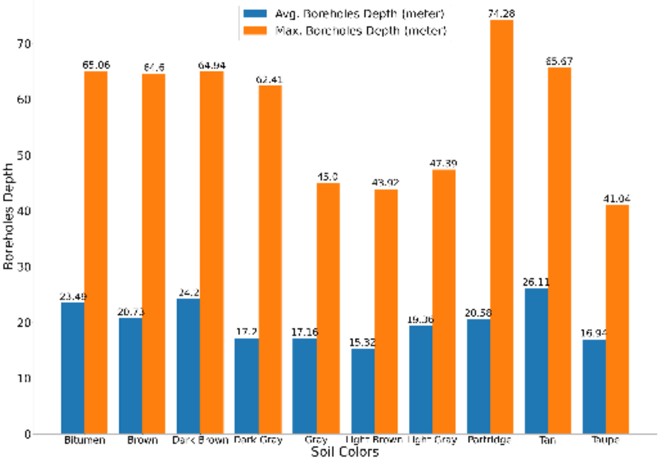


FIGURE 7. Soil color analysis based on boreholes depth.

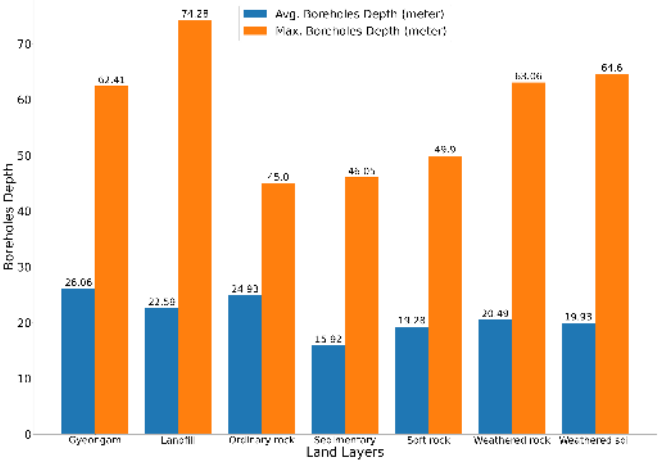


FIGURE 8. Land (Rock) layer analysis based on boreholes depth.

Similarly, Fig. 8 analyzes boreholes data based on land layer according to an average and maximum boreholes depth during drilling to gain water levels. The analysis results reveal that the land layer "Gyeongam" has maximum average boreholes depth of 74.28 meters, and "Sedimentary" layer has minimum average boreholes depth of 15.92 meters. Likewise, landfill layer has maximum boreholes depth of 74.28 meters, which shows that the drilling process took a large amount of time to drill under the earth’s surface to reach the water levels. Hence, drilling through ordinary and soft rock units is easier and time-saving than the other land layers to gain the GWL.

C. FEATURES EXTRACTION

Features extraction is a vital process to construct new features based on the existing data features. It also reduces dimensionality. The feature extraction techniques aim to enhance model accuracy, overcome overfitting issues, speed up model training, and reduce computational complexity. In this study, some of new features are computed using existing data attributes, such as total depth of boreholes drilling, days spent on each borehole drilling, core soil color, core stratum layer, and core land layer.

Borehole depth is defined as the sum of the thickness (T) of the land layer for each borehole log. Thickness is determined by taking the difference between the top (starting) and bottom (ending) drilling depth of each land layer. Thickness is calculated as shown in equations 1 and 2.

$$T = ED - SD \tag{1}$$

$$T = \sum_{i \in j}^{M} (ED_i - SD_i) \tag{2}$$

The total boreholes depth (TB_{depth}) is calculated as the sum of the thickness instances for each borehole log location i. (TB_{depth}) is computed as shown in equation 3.

$$TB_{depth} = \sum_{i \in B}^{N} \sum_{j \in i}^{M} T_i \tag{3}$$

Fig. 9 shows analysis of boreholes log data based on drilling depth of boreholes log and days spent on drilling to gain the groundwater level. The analysis investigated the relationship between total boreholes depth and drilling time for each unique borehole location. According to the analysis

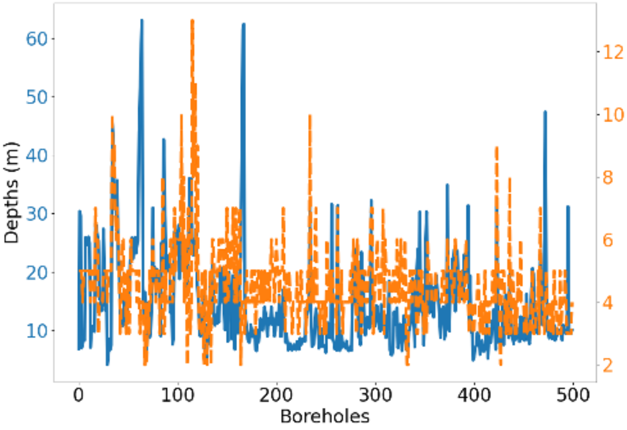


FIGURE 9. Borehole log data analysis based on drilling depths and time duration.

results, most of the time, it is found that drilling time is minimum and boreholes depth is maximum, which indicates that the drilling process is easier to gain GWL in the selected regions.

Next, a temporal feature is calculated to analyze the total number of days spent on the borehole to reach the GWL. The total number of days is defined as counting the combinations of thickness the rock units for each drilling location. Equation 5 is used to compute the time duration (TD) for each drilling location.

$$TD_{days} = Count\ Thickness\ Instances \tag{4}$$

Furthermore, box-plot analysis is widely used to measure five value summary, such as minimum, lower quartile of the median, median, upper quartile of the median, and maximum values. Fig. 10 shows box-plot analysis to investigate GWL according to time interval groups (in terms of days). It can be seen that the relationship between TD_{days} and GWL varies because of the different structures of rock units. As an example, it can be observed that 5 to 6 days spent to gain GWL between 0.35 m to 22.2 m. Data outliers are figured out that are distant from the scattered data samples. The data points visualized outside of the box-plot whiskers are defined as data outliers. Furthermore, in the case of 11 to 13 days, it can be analyzed that GWL varies between 2.8 m and 7.09 m. Moreover, hardness of rock layers ultimately minimizes GWL and maximizes time spent.

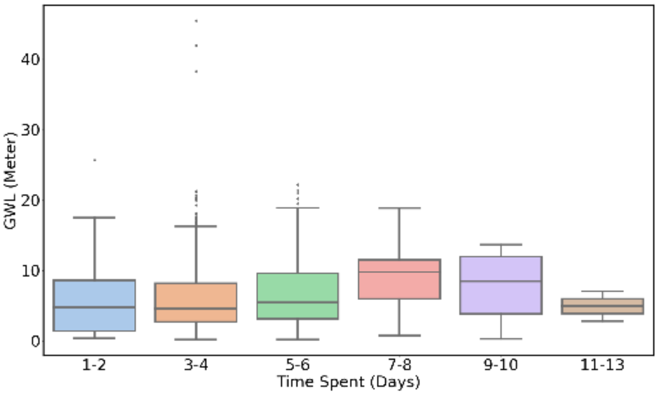


FIGURE 10. GWL analysis according to time taken intervals.

Fig. 11 analyze drilling data based on TB_{depth} and GWL according to the days spent on each drilling location. It can be seen that the relationship between all these three attributes is varied due to the different structures of rock layers. The resulting analysis shows that days spent on each drilling location ranges from 1 to 13 to reach the GWL. Similarly, GWL fluctuates between 0.17 m to 45.5 m to extract water in the scenario area. Besides, the drilling depth of the boreholes log is up to 74.2 m to access the GWL in the boreholes region. The analysis results depict that an average of days spent on each drilling location is 5 to access the GWL. Furthermore, it can be examined that the drilling depth of boreholes and GWL varies because of the different structures of rock and soil patterns, which also influences the time taken by each borehole to drill.

The next feature is core soil color, which is extracted based on maximum total boreholes depth. The drilling for each borehole log consists of different soil colors patterns. Algorithm 1 presents a detailed flow of the core soil color for each borehole. The boreholes data and unique boreholes are used as input data. The objective of the algorithm is to extract the core soil color based on maximum drilling depth for each unique borehole. It is earlier discussed that the drilling process for each borehole consists of several soil colors. Therefore, first of all, unique soil colors are extracted for each borehole. Second, total drilling depth is calculated for each unique soil color. Finally, a soil color with maximum drilling depth denoted as a core soil pattern for an i^th drilling location.

Algorithm 1: Extraction of core soil color for each unique borehole.

```
Input: Input Boreholes Data Samples $B_{data}$ , unique
       boreholes $u$
Output: Core soil color for each unique borehole
        $Core_{soilcolor}$
$u \leftarrow uniqueBoreholes(B_{data})$
for $i \in u$ do
    $Soil_{colors} \leftarrow uniqueSoilColors(i)$
    $Borehole_{depth} \leftarrow 0.0$
    $Core_{color} \leftarrow null$
    for $s \in Soil_{colors}$ do
        $B_{depth} \leftarrow depth(i)$ ;   // Get borehole
         depth according to soil color
         $s$ for $i_{th}$ borehole location
        if $B_{depth} > Borehole_{depth}$ then
            $Borehole_{depth} \leftarrow B_{depth}$
            $Core_{color} \leftarrow s$ ;     // Assign soil
             color $s$ with the maximum
             depth to $Core_{color}$
        end
    end
    $Core_{soilcolor}[i] \leftarrow Core_{color}$
end
```

The extraction flow of the core land layer and stratup layer for each borehole is given in algorithm 2. The drilling process of the boreholes consists of several land layers and stratup layers to reach the GWL. Therefore, it is needed to analyze and find the core land and stratup layers based on maximum borehole depths. Hence, a core layer is defined as the land layer with maximum drilling depth for an i^th borehole. Similarly, a core stratup layer is defined as the stratup layer with a maximum frequency of drilling for an i^th borehole. Therefore, for each unique borehole, a drilling frequency for each unique land layer is computed to analyze and select a land layer as a core land layer having maximum drilling frequency. Similarly, according to the stratup layers, a drilling frequency is also computed to analyze and find a core stratup layer with maximum boreholes depth.

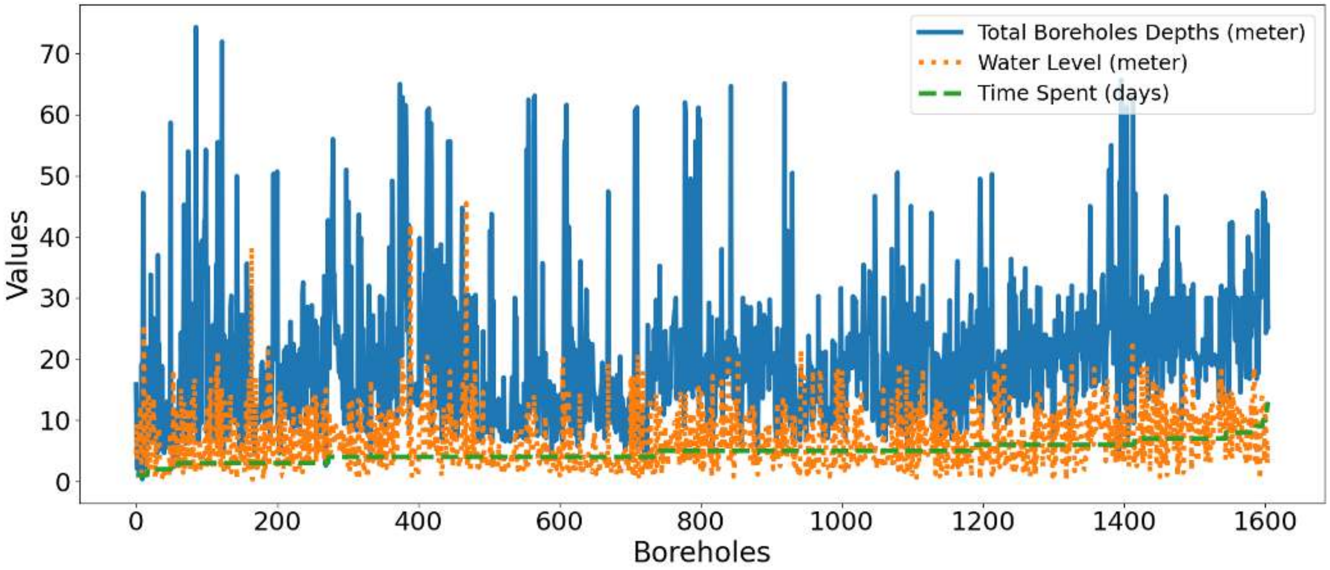


FIGURE 11. Borehole depth and GWL analysis according to time spent.

Algorithm 2: Extraction of core land and stratup layer for each unique borehole.

```
Input: Input Boreholes Data Samples $B_{data}$ , unique
       boreholes $u$
Output: Core land layer and Core Stratup layer for
        each unique borehole $Core_{landlayer}$ and
        $Core_{stratuplayer}$
$u \leftarrow uniqueBoreholes(B_{data})$
for $i \in u$ do
    $Land_{layers} \leftarrow uniqueLandLayers(i)$
    $Stratup_{layers} \leftarrow uniqueStratupLayers(i)$
    $Borehole_{depth} \leftarrow 0.0$
    $Core_{land} \leftarrow null$
    $Core_{stratup} \leftarrow null$
    for $l \in Land_{layers}$ do
        $B_{depth} \leftarrow depth(i)$
        if $B_{depth} > Borehole_{depth}$ then
            $Borehole_{depth} \leftarrow B_{depth}$
            $Core_{land} \leftarrow l$
        end
    end
    $Core_{landlayer}[i] \leftarrow Core_{land}$
    $Borehole_{depth} \leftarrow 0.0$
    for $s \in Stratup_{layers}$ do
        $B_{depth} \leftarrow depth(i)$
        if $B_{depth} > Borehole_{depth}$ then
            $Borehole_{depth} \leftarrow B_{depth}$
            $Core_{stratup} \leftarrow s$
        end
    end
    $Core_{stratuplayer}[i] \leftarrow Core_{stratup}$
end
```

D. FEATURES NORMALIZATION AND SELECTION

This subsection describes features normalization and selection. Data normalization is an important process to scale down feature values in some specified range, for instance, [0,1]. It is an effective process to transform data into a common scale to avoid biases among data features and improve model learning. Therefore, a feature normalization is required because the range of feature values is different. Different features normalization techniques are considered, for instance, min-max normalization, z-score, clipping, etc. This research study utilizes min-max normalization to scale down feature values in a similar range to consider each feature equally in the model learning process.

$$\hat{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

The next step is to select the most promising features to reduce the high dimensionality of the dataset and improve the performance of the model without losing information. Commonly used feature selection techniques are correlation analysis, information gain, principal component analysis, to name a few. This work uses correlation analysis as a benchmark technique to compute the correlation index of all input features with respect to the target feature and select those features having a correlation index 03.0 or greater than 0.30. The correlation heatmap map is shown in Fig. 12 to analyze a linear relationship between input features and target features.

It can be observed that altitude and temporal difference features are negatively correlated with a target feature; therefore, both features are removed from the given feature space to reduce the computation and storage cost. Fig. 13 presents a correlation heatmap for the selected features to analyze the linear relationship between independent (soil color, total

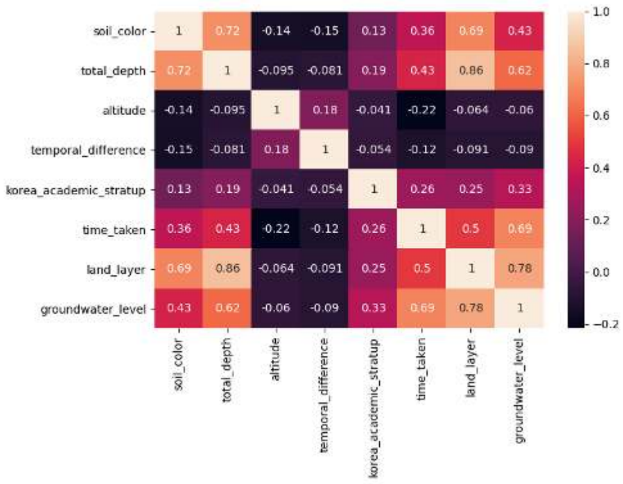


FIGURE 12. Features selection based on correlation analysis.

depth, stratum layer, time taken, and land layer) and dependent variables (groundwater level).

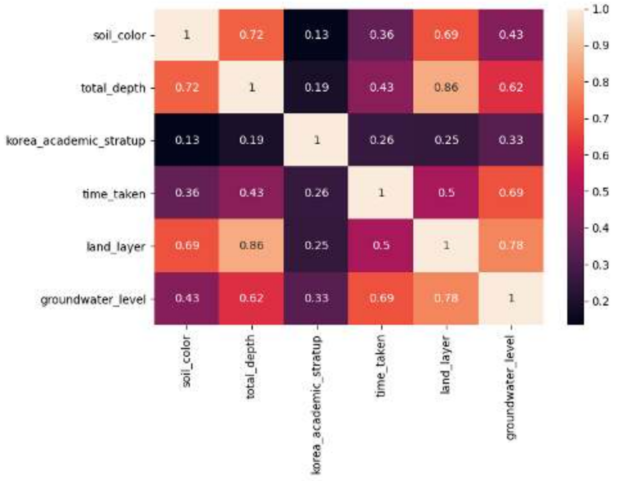


FIGURE 13. Selected features set.

VI. PROPOSED DIFFERENCE MECHANISM

This section presents difference model and the implementation of drilling dataset. It also includes the implementation of the following functions, such as input function, comparison function, and search function for drilling dataset. Furthermore, it suggests a logical/mathematical model for enhancing data search results.

Time series data can be transformed using a technique called differencing to eliminate temporal dependence. Before modelling time-series data, the trends and seasonality factor might need to be removed. To achieve this difference is utilized as an effective data transformation method for constructing stationary time series data. For statistical modelling techniques, time series should be stationary for ease in modelling. As non-stationary time series data possess specific trends and seasonality that vary with time. Likewise, the statistical measures incur changes with time, for example, mean, and variance, which leads to change in concept which model is trying to learn.

For the transformation of a time-series dataset, various differencing methods are utilized. Differencing methods are an effective way to eliminate temporal dependence that exists in a time series, more specifically concerning features related to trend and seasonality in data. Moreover, it can remove variations in time series by achieving a stable mean and ultimately lessens the impacts of trends in data. It works by computing the difference between current and previous data sample values.

Differencing measure involves methods that compare two time-series objects and output a value that encodes how dissimilar they are. The distance can be defined as a quantitative measurement of dissimilarity or difference, specifying how far two instances are from each other. Fig. 14 presents difference between consecutive starting borehole depth samples using lag-1 difference. It can be observed that average starting borehole depth rate varied between 3 meter to 25 meter.

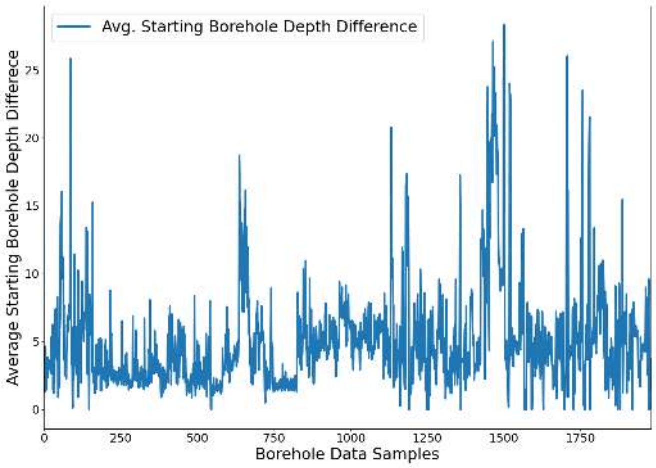


FIGURE 14. Starting boreholes depths analysis using lag-1 difference algorithm.

Similarly, Fig. 15 is used to presents an average temporal difference of ending borehole depth for each borehole location. The ending depth of boreholes locations indicates the bottom part of the rock unit during thickness combination. The difference between the top and bottom frequency of drilling depth is defined thickness, which fluctuates due to different hydrogeological patterns and climatic changes. It can be observed that average ending borehole depth rate varied between 3 meter to 45 meter.

VII. IMPLEMENTATION AND EXPERIMENTAL SETUP

In this section, an experimental setup of the proposed E-GWLP model is presented. In this work, we used Python as a core language to implement and conduct a series of

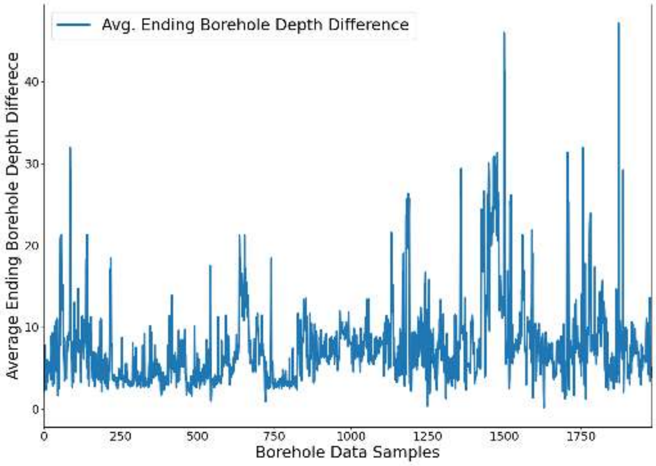


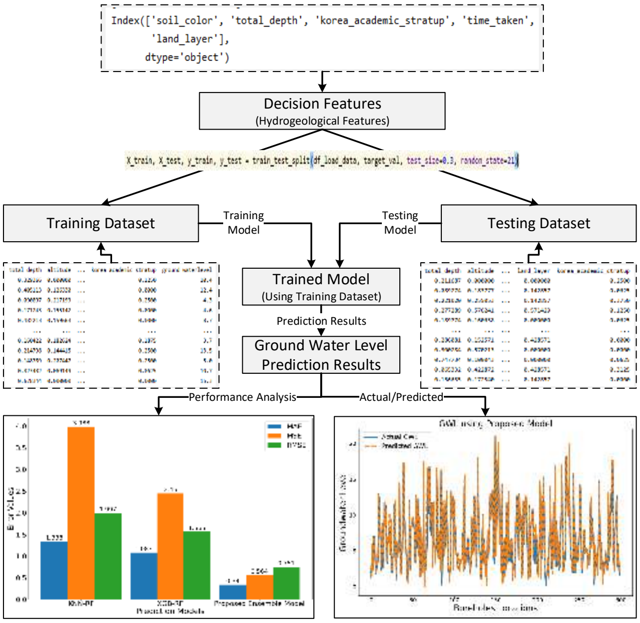
FIGURE 15. Ending boreholes depths analysis using lag-1 difference algorithm.

experiments. The following core libraries of Python are utilized, such as Sklearn, Seaborn, Matplotlib, Pandas, Numpy, to name a few. Furthermore, MS Excel and MySQL are used to store raw and process boreholes data. Moreover, we used Intel Core i9-10900 CPU along with 32 GB RAM to perform experiments. Table 2 summarizes the experimental setup of the E-GWLP model.

TABLE 2. Implementation and experimental setup of the proposed ensemble model.

Components	Description
Operating System	Microsoft windows 10 (64-bits)
Processor	Intel ®Core ™ i9-10900 CPU at 2.80 GHz
Main Memory	32 GB
Backend Language	Python
Storage	MS Excel and MySQL
IDE	PyCharm Professional
Core Libraries	Sklearn, Seaborn, Matplotlib, Numpy, and Pandas.

Figure 16 depicts implementation process of the proposed E-GWLP model. Our proposed E-GWLP model utilized python as the core backend programming language to performed different experiments, including data and predictive analysis. A sklearn library is used to utilize various features, such as the transformation of categorical values into continuous values, division of prepared boreholes data into training and testing samples sets, training and testing of ML-based regression models. Min-max scaler is used to mapped the feature values into a specified range [0,1] to overcome the learning issues of ML models. The prepared dataset is divided into training samples and testing samples with a split ratio of 70-30; it indicates that 70% of boreholes samples are used for building ML models, and the remaining 30% of boreholes samples are utilized for evaluation purpose. Furthermore, different evaluation measurements are considered to evaluate the error of the each regression model.



```
Index(['soil_color', 'total_depth', 'korea_academic_stratup', 'time_taken',
       'land_layer'],
      dtype='object')
```

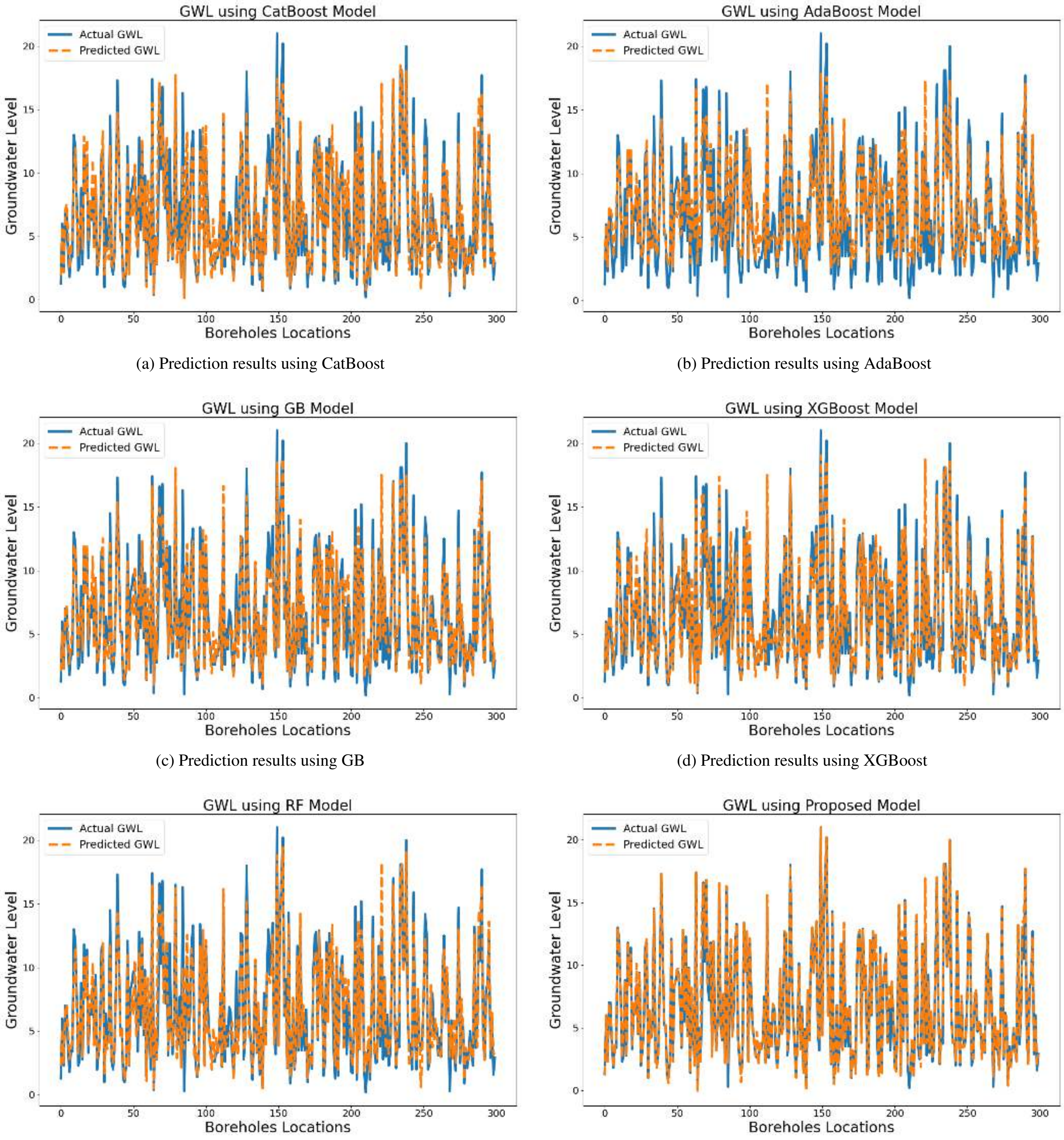
```
X_train, X_test, y_train, y_test = train_test_split(df_load_data, target_val, test_size=0.3, random_state=21)
```

FIGURE 16. Implementation flow of the proposed E-GWLP model.

VIII. IMPLEMENTATION RESULTS AND ANALYSIS

This section provides a detailed overview of the results yield by experiments moreover a detailed performance analysis is also presented for GWL prediction. There are two types of experimental results analyses performed. First, prediction results of our E-GWLP model is compared with traditional ensemble model to highlights the significance of the proposed work. Second, experimental results of our model is compared with baseline regression models.

Fig. 17 depicts a comparison of the implemented regression models to predict GWL. In Fig. 17 the observed and estimated GWL are analyzed. The analysis verify that the proposed framework based on the ensemble model outperformed conventional methods. Fig. 20a presents the actual vs predicted GWL based on the CatBoost model. The difference analysis of actual and forecasted is justifiable comparative to Adaboost and GB. Similarly, Fig. 17b depicts the prediction error of the AdaBoost model, It can be seen that the gap between actual and forecasted values is comparatively high than those achieved by CatBoost and GB models. In Fig. 17c, showcase a comparison of actual and predicted values using the GB model. It can be seen clearly that the prediction error is relatively higher compared to the CatBoost model. Furthermore, in Fig. 17d, it is evident that the prediction error of the XGBoost model is low compared to the CatBoost, AdaBoost, and GB models. Fig. 17e shows a comparative analysis of actual GWL and predicted acquired by RF, which indicates that RF produced slightly high error compared to the XGBoost and CatBoost models. Finally, Fig. 17f visualized actual and predicted GWL using the proposed ensemble model. It can be clearly seen that occurrence of prediction error by using proposed ensemble model is lower comparative



(a) Prediction results using CatBoost

(b) Prediction results using AdaBoost

(c) Prediction results using GB

(d) Prediction results using XGBoost

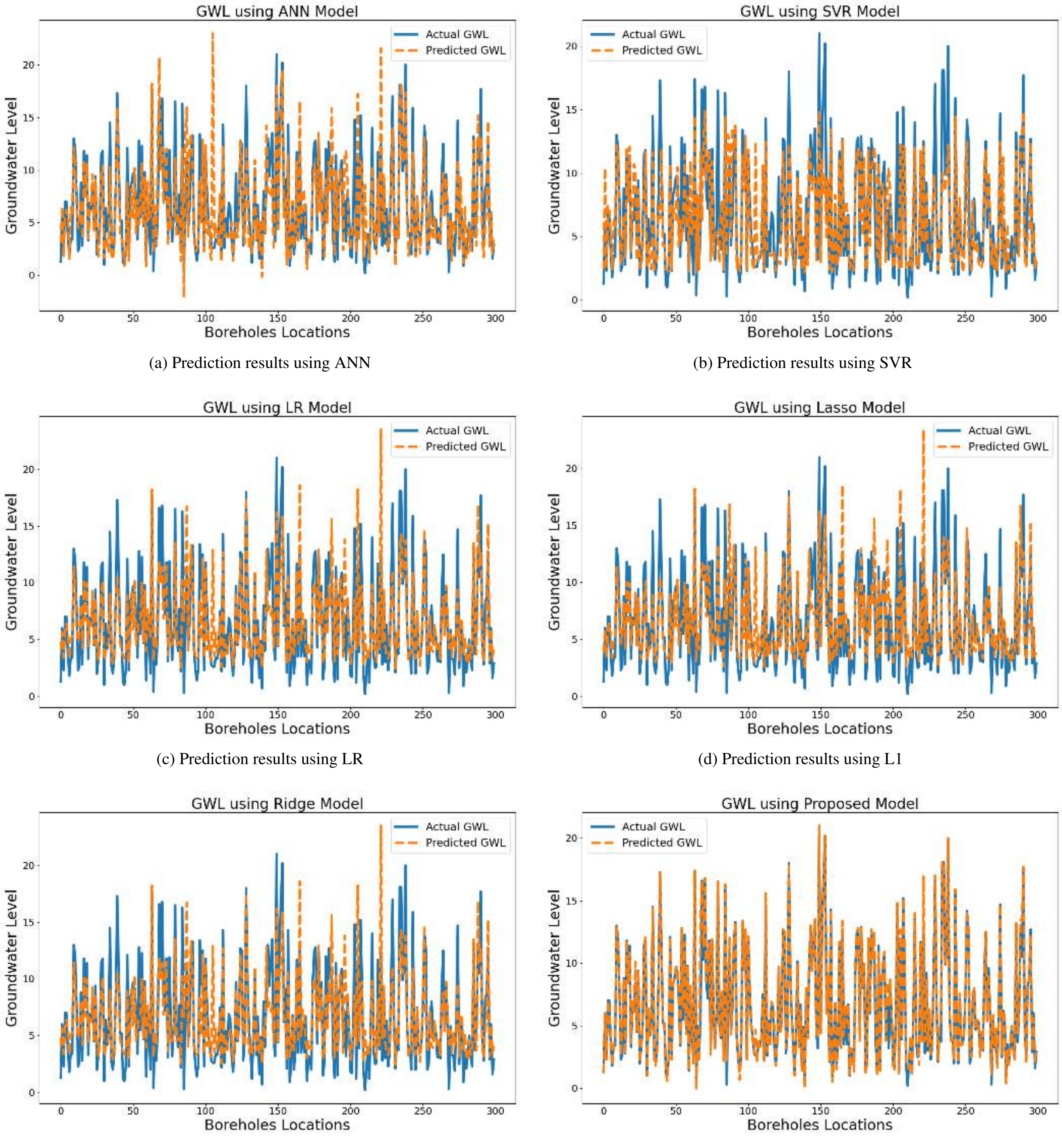
(e) Prediction results using RF

(f) Prediction results of E-GWLP model

FIGURE 17. Comparison of E-GWLP model with traditional ensemble approaches for GWL predictions.

to counterpart solutions, including GB, CatBoost, AdaBoost etc. This verify the proposition of the study that proposed ensemble prediction model yield superior performance by achieving a low prediction error and can be considered a sustainable solution for enhancing future boreholes efficiency and reservoir engineering.

Furthermore, Fig. 18 visualizes actual and predicted values achieved by proposed ensemble model along with its comparison with baseline regression models. Fig. 20b shows that the conventional ANN model produced a relatively high



(a) Prediction results using ANN

(b) Prediction results using SVR

(c) Prediction results using LR

(d) Prediction results using L1

(e) Prediction results using L2

(f) Prediction results of E-GWLP model

FIGURE 18. Comparison of E-GWLP model with traditional learning models for GWL predictions.

prediction error compared to the proposed ensemble model. Similarly, Fig. 18b indicates the prediction results of the baseline SVR model, it is evident from the comparison that our model achieved lower error percentage compared to the ANN and LR. Moreover LR model also produced a high prediction error compared to ANN and SVR models as shown in Fig. 18c. Similarly Fig. 18d and Fig. 18e analyzes prediction error for unseen data samples using L1 and L2 models. It can be observed that prediction error in case of using L1 and L2 models are higher comparative to ANN and SVR models.

The prediction error observed in case of conventional models for unseen samples is significantly high and cannot be considered those models to predict GWL for future boreholes. The comparative review reveals that the prediction results of the conventional statistical models are not acceptable for sustainable water resource management. Hence, it can be concluded that our proposed ensemble model has achieved satisfactory results and outperformed the conventional regression models by bringing massive improvements concerning performance of the GWL prediction. The findings of experimental results prove that proposed ensemble model is suitable for predicting GWL to enhance the efficiency of future water boreholes.

Furthermore, Fig. 19 shows the comparison of proposed framework with hybrid prediction models. The analysis reveals that prediction error caused by KNN-RF and XGB-RF models is slightly higher in comparison to our proposed E-GWLP model. Hence, our proposed ensemble model produced more accurate results in contrast to aggregated mean-based hybrid prediction framework.

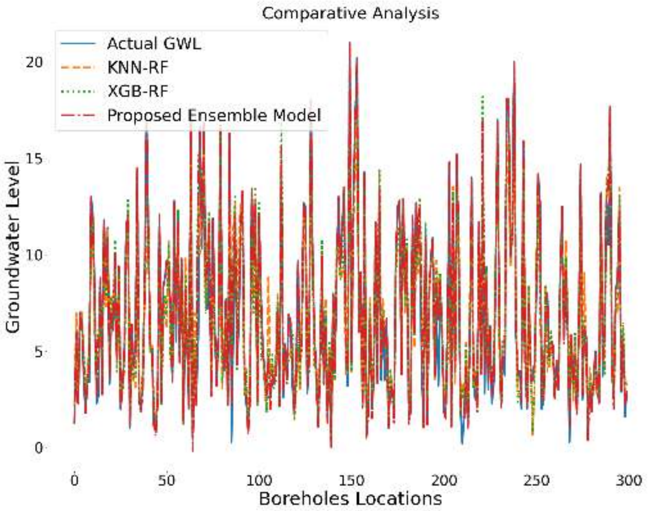


FIGURE 19. Comparative analysis of the proposed ensemble models

Features importance is an important process to investigate the significance of the prepared data features [61]. Feature importance refers to assigning importance to the feature variables based on specific scores. Scores are allocated based on their usefulness at predicting the output variable. It can be used for dimensionality reduction by selecting only promising features from the given feature space. Faster training and complexity reduction, and easy interpretation are some advantages of applying feature importance. Furthermore, it is an efficient way to find the contribution of each feature in the model learning phase and eliminate the least contributed features from the features space to produce generalize and accurate decision model. Therefore, it is required to identify the most contributed features in the prepared dataset. Figure 20 shows a comparison of features importance using conventional ensemble models. XGBoost indicates that the temporal difference feature has a highly contributed feature compared to other proposed features. Adaboost, RF, and GB models indicate that the score of the altitude feature has high, which means that the altitude feature contributed more compared to the other listed features.

The proposed study employed various statistical formulations for measuring the forecasting error of conventional ensemble models and baseline ML models. Performance analysis metrics include widely used metrics including MAE, MSE, RMSE, normalized RMSE (NRMSE), MAPE, and R2 scores. MAE and MSE are common performance evaluation measure used for continuous variables [53], [62].

MAE measures the difference between actual and estimated values by extracting the average of absolute difference based on entire dataset and provides the average error magnitude. It is formulated as shown in 6:

$$MAE = \sum_{i=1}^{n} |y_{observed} - y_{estimated}| \tag{6}$$

MSE measures the difference between estimated and actual values (residuals) and resultingly provide a value that depicts how closely the fitted lined lies to the data points, lastly the value is squared so that negative values turn positive. the lesser the value of mean square error the closer the fit, better is model performance. It is obtained by finding the difference then taking average of squared value and calculating square root finally. MSE is calculated using the following equation 7.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{observed} - y_{estimated})^2 \tag{7}$$

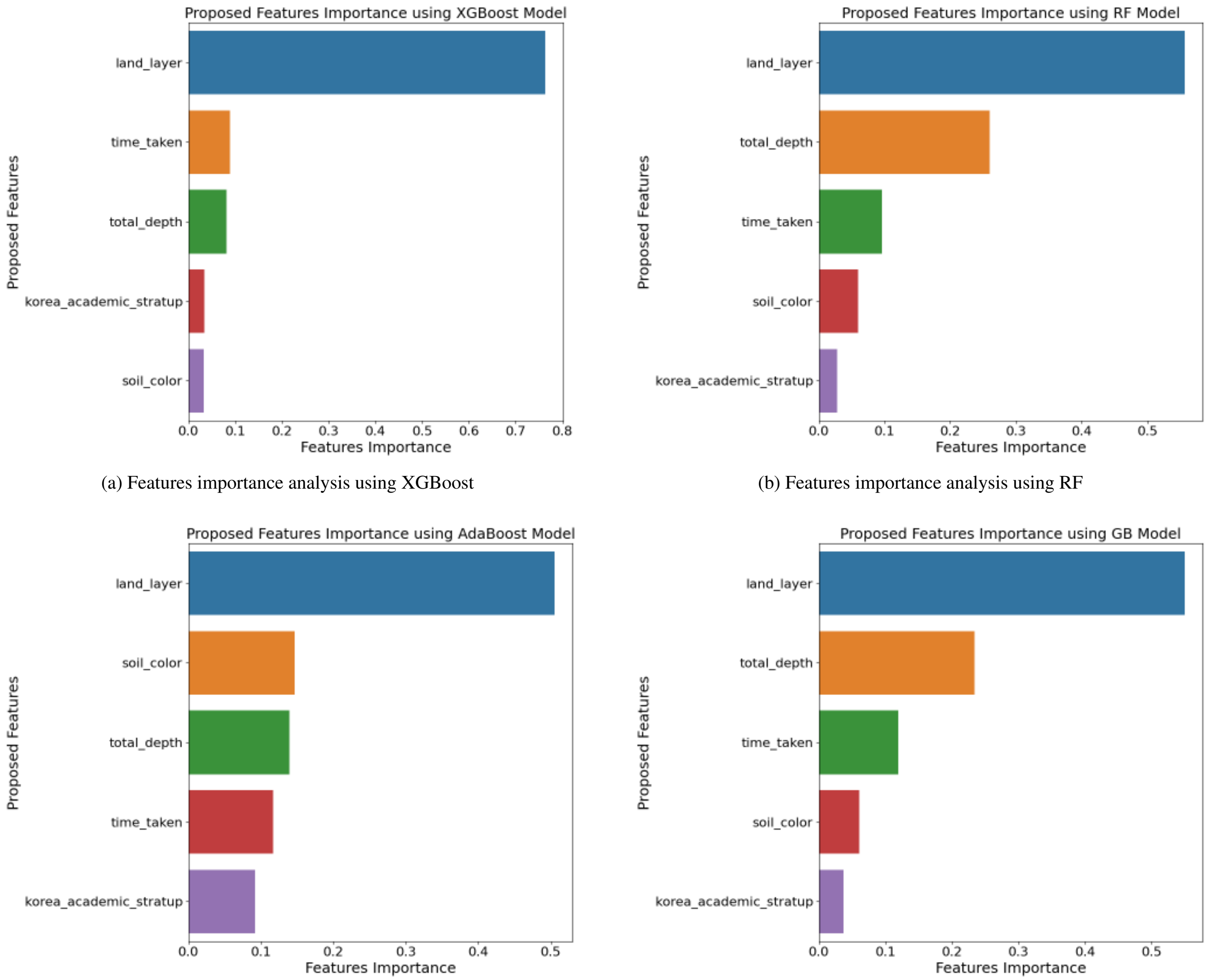
RMSE is a defined as the square root of the MSE. It is used to measure the average distance that starts from fitted line to the data points along vertical axis. The formula for calculating RMSE is provided in equation 8.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{observed} - y_{estimated})^2}{n}} \tag{8}$$

R2 score is a statistical measure that is defined as a coefficient of determination that involves observed and predicted values for evaluating how well the regression model performs. R2 score approaching 1 or close to 1 is an indicator of good performance achieved by regression model. R2 score is computed based on the following equation 9.

$$R^2\ Score = 1 - \frac{\sum(y_{observed} - y_{estimated})^2}{\sum(y_{observed} - \bar{y}_{estimated})^2} \tag{9}$$

MAPE is another statistical measure to estimate the regression model's accuracy in terms of differences between observed and predicted values. It is defined as an average of the absolute percentage errors of the regression model. The low MAPE indicates the high accuracy percentage of the prediction model. The basic formula is given in equation 10 to measure MAPE.



(a) Features importance analysis using XGBoost

(b) Features importance analysis using RF

(c) Features importance analysis using AdaBoost

(d) Features importance analysis using GB

FIGURE 20. Proposed features importance analysis using XGBoost, RF, AdaBoost, and GB models.

$$MAPE = \frac{1}{n}\sum_{k=1}^{n}\frac{(y_{observed} - \hat{y}_{estimated})}{y_{observed}} \quad (10)$$

Table 3 presents performance evaluation of the proposed model along with comparative analysis with counterpart conventional learning models, including CatBosot, AdaBoost, GB, XGBoost, and RF. Furthermore the proposed model's performance is compared with some developed integrated models. These models include KNN-RF and XGB-RF. The experimental findings made the fact evident that validation and testing performance of the proposed model is superior than all other standalone and ensemble models. The validation performance analysis also proves model's strength as it successfully achieves a lower MAPE value and high R2 score compared to the baseline models. In the validation analysis, our proposed ensemble prediction model gained MAPE of 13.473 and R2 score of 0.945. Similarly, the testing results also proved that ensemble prediction framework proposed in this study produced accurate results comparative to counterpart solutions. The testing performance of proposed model reported a MAPE 12.658 and R2 score of 0.976. Furthermore, MAE, MSE, RMSE produced by the proposed solution is 0.340, 0.564, and 0.751, respectively. The experimental findings proves the efficiency and robustness of our proposed model compared to the conventional bagging and boosting models. The NRMSE value reported by our model is 0.018. The scores achieved by the proposed method highlights the significance of the our proposed model. On the contrary MAPE produced by CatBosot, AdaBoost, GB, XGBoost, and RF models is 24.394, 47.079, 31.014, 26.647, and 29.146, respectively. Furthermore, the results of proposed study are also compared to hybrid models including; KNN-RF and

XGB-RF. It can be observed that the MAPE of KNN-RF and XGB-RF is higher compared to our proposed ensemble model. It is proven that the forecasting error of unseen samples using proposed E-GWLP model is significantly low than conventional bagging and boosting models. Hence, based on performance analysis results, it is proved that our E-GWLP model performed significantly better to predict GWL compared to the conventional models and state-of-the-art techniques.

Fig.21 depicts MAE, MSE and RMSE error values to evaluate the forecasting accuracy of our model compared to traditional aggregated-mean based ensemble models. The comparative analysis shows that the forecasting error of the proposed model is low compared to the KNN-RF and XGB-RF models. The evaluation analysis indicates that estimation error of the XGB-RF model is slightly low then the traditional KNN-RF model. Overall, our proposed ensemble model performed precisely and correctly comparative to conventional ensemble modeling based solutions. The MSE and RMSE error values of the proposed ensemble model are 9.735 and 3.12, respectively, which proved that the proposed ensemble prediction model accurately predicts GWL for enhancing management of hydraulic resources. In contrast, MSE error values of the conventional ensemble model are high; it can be clearly seen that the MSE values of the BME, AdaBoost, and GB are 18.449, 15.803, and 11.246, respectively. Hence, our proposed ensemble solution is robust and accurate in predicting GWL compared to the conventional ensemble models.

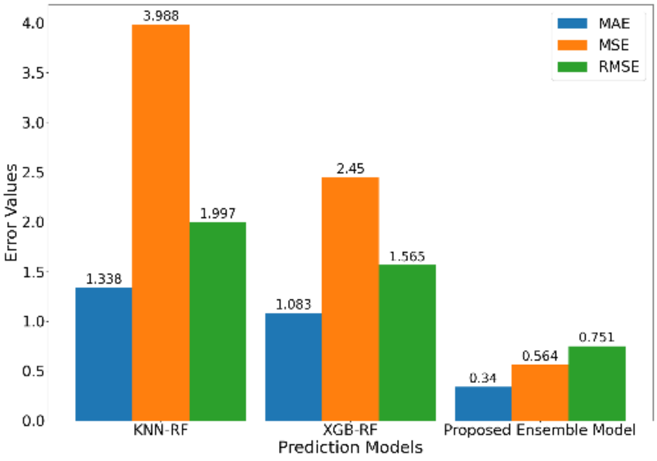


FIGURE 21. Evaluation analysis of the prediction error.

Furthermore, Fig. 22 illustrate performance evaluation of proposed model comparative to conventional prediction model. The MAPE metric is considered to evaluate the prediction accuracy of the proposed study and other standalone and ensemble models. Our model observed a MAPE of 12.66, that is an indicator of effectiveness of our proposed solution comparative to regression models. The MAPE values of the conventional regression models, including SVR, ANN, LR, and DT, are 4240, 47.2, 53.38, 24.53, respectively. Similarly, MAPE values of the conventional ensemble models, including KNN-RF and XGB-RF, are 34.53 and 28.44, respectively. The evaluation analysis shows that the DT-based regression model performed relatively better than other conventional and ensemble models. The DT yield a low error in terms of MAPE of 24.531. Lastly our E-GWLP model produced lowest MAPE compared to counterpart solution that verify the efficiency of our proposed solution.

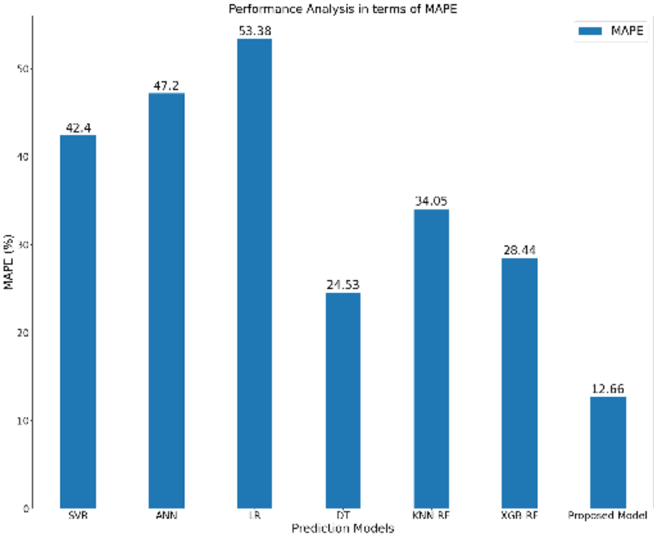


FIGURE 22. Performance evaluation of the prediction models based on MAPE.

Table 4 compares evaluation results of our proposed ensemble model with baseline learning models; ANN, SVR, LR, L1, and L2. The validation error produced by our model in terms of MAPE is 12.658, observed to be lowest among all baseline regression models. On the other hand MAPE (validation performance) of baseline models, including ANN, SVR, LR, L1, and L2, is 42.399, 40.845, 49.650, 53.664, and 53.377 respectively. Furthermore, the observed MAPE value produced by our proposed solution for test instances is 12.658 , that is an indicator of our models predictive power comparative to conventional solutions. In contrast, MAPE value of the conventional ANN, SVR, LR, L1, and L2 models is 47.202, 42.399, 53.377, 49.379, and 49.650, respectively, which shows that traditional models are performed poorly compared to our proposed ensemble model. The L1 model performed slightly well than the baseline models; MAE and MSE value of L1 is 1.693 and 5.991, respectively. The performance analysis of proposed model established the fact that error rate achieved by our model is significantly low comparative to baseline solution approaches. In contrast, the RMSE value of the ANN, SVR, LR, L1, and L2 values are 2.87, 3.089, 2.454, 2.395, and 2.394, respectively. Hence, it is proved that our E-GWLP model performed accurately and precisely to predict GWL compared to baseline prediction models.

Furthermore, in Fig. 23,we presented an evaluation analysis involving proposed solution and conventional regression

TABLE 3. Performance analysis and comparison of the proposed E-GWLP model with conventional ensemble models.

Models	Validation Performance (k-fold)						Testing Performance (Unseen Samples)					
	MAE	MSE	RMSE	NRMSE	MAPE	R2 Score	MAE	MSE	RMSE	NRMSE	MAPE	R2 Score
CatBoost	1.185	3.694	1.872	0.082	24.605	0.835	1.247	3.710	1.926	0.046	24.394	0.844
AdaBoost	1.499	4.225	2.031	0.098	45.598	0.809	1.464	3.261	1.806	0.046	47.079	0.863
GB	1.171	3.096	1.741	0.080	28.074	0.856	1.211	2.708	1.646	0.040	31.014	0.886
XGBoost	1.096	2.867	1.688	0.076	23.839	0.866	1.113	2.564	1.601	0.039	26.647	0.892
RF	1.105	3.270	1.784	0.080	24.906	0.850	1.092	2.538	1.593	0.038	29.146	0.893
KNN-RF	1.449	4.207	1.534	0.072	26.912	0.823	1.338	3.988	1.997	0.048	34.048	0.832
XGB-RF	1.121	2.630	1.698	0.068	22.263	0.867	1.083	2.450	1.565	0.038	28.444	0.897
Proposed Model	**0.562**	**0.723**	**0.821**	**0.021**	**13.473**	**0.945**	**0.340**	**0.564**	**0.751**	**0.018**	**12.658**	**0.976**

TABLE 4. Performance analysis and comparison of the proposed E-GWLP model with conventional ML models.

Models	Validation Performance (k-fold)						Testing Performance (Unseen Samples)					
	MAE	MSE	RMSE	NRMSE	MAPE	R2 Score	MAE	MSE	RMSE	NRMSE	MAPE	R2 Score
ANN	1.247	3.710	1.926	0.046	42.399	0.599	1.793	8.238	2.870	0.065	47.202	0.654
SVR	1.830	8.505	2.888	0.138	40.845	0.614	1.857	9.545	3.089	0.078	42.399	0.599
DT	1.354	5.378	2.282	0.102	27.547	0.754	1.220	3.785	1.946	0.047	24.531	0.841
LR	1.695	6.003	2.394	0.115	49.650	0.732	1.804	6.020	2.454	0.062	53.377	0.747
L1	1.806	6.049	2.459	0.062	53.664	0.746	1.693	5.991	2.395	0.116	49.379	0.732
L2	1.804	6.020	2.454	0.062	53.377	0.747	1.695	6.003	2.394	0.115	49.650	0.732
Proposed Model	**0.562**	**0.723**	**0.821**	**0.021**	**13.473**	**0.945**	**0.340**	**0.564**	**0.751**	**0.018**	**12.658**	**0.976**

based model solution approaches.

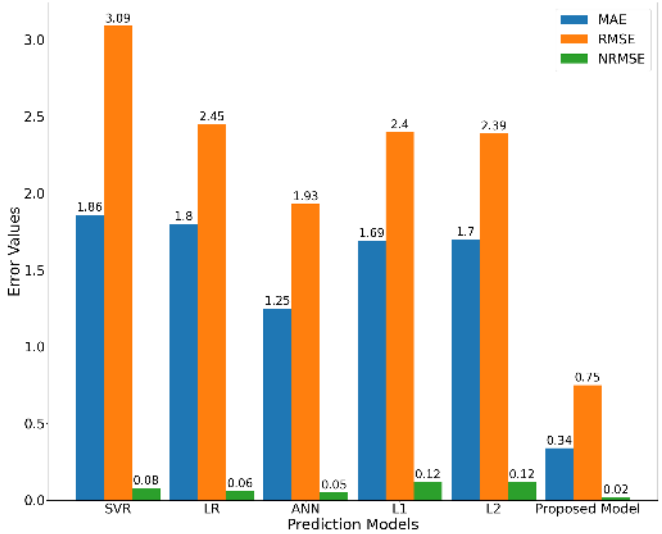


FIGURE 23. Comparison of the XGB-RF with baseline regression models.

MAE, MSE, and N-RMSE error metrics are considered to analyze the prediction error of the proposed ensemble and traditional regression algorithms. The MAE, MSE, and NRMSE values of E-GWLP approach are 0.340, 0.564, and 0.018, respectively. The analysis revealed that high error rates are produced by conventional ANN and linear regression (LR) models are high in comparison to baseline models that include SVR, lasso (L1), and ridge (L2). Results are an indicator of how well our proposed E-GWLP model generalized the data, and produced accurate prediction results.

Moreover, Table 5 shows a comparative analysis of our E-GWLP and existing state-of-art models. Different important parameters are taken into account to compare the proposed study results with the existing model. The comparative analysis results indicates that the existing model used a traditional approach to combine KNN and RF using aggregated mean to form an ensemble model. However, our proposed model is developed based on the stacking technique to forecast GWL. Furthermore, an existing model used the following input features, including temperature, precipitation, and solar radiation, to predict GWL. In contrast, our proposed ensemble model used hydrogeological and time interval features to forecast GWL to improve hydraulic management. It can also be observed that the baseline model used a sliding window-based approach to validate trained models. In comparison, our work used the k-fold validation method for validating models to avoid overfitting issues. Moreover our proposed model produced r2 scores of 0.976, contrarily R2 score observed in case of existing model is 0.939. Hence, our proposed E-GWLP model is a reliable solution compared to the existing prediction model to improve hydraulic resource management effectively.

Moreover, Fig. 24 presents features importance of the proposed ensemble model. The impact of hydrogeological and time interval features are analyzed with respect to their impact towards the GWL prediction. Information gain (IG) is known because of its widespread use for determining how impactful any each feature has on prediction process. We analyzed the features importance discovered that the core land layer has the most promising feature with high IG. The total time spent feature is the second important feature among selected features. Furthermore, total borehole depth also has high IG compared to the core stratum layer and core soil color features. Moreover, core stratum layer has a low

TABLE 5. Comparison of the proposed E-GWLP model with existing prediction model.

Parameters	Proposed Ensemble Model	KNN-RF [39]
Prediction Models	XGBoost, AdaBoost, and GB (Level 0) and RF (Level 1)	KNN and RF
Combinator	Stacking	Aggregated Mean
Total Features	5	3
Data Features	Core land layer, Core soil color, Core Stratum Layer, Total time taken, and Total borehole depth	Temperature, Precipitation and Solar radiation
Validation Scheme	10-Fold cross-validation	4-Sliding window based validation
R2 Score	0.976	0.939

IG of 0.4 among the considered features set. The features importance analysis reveals that the rock layer feature has been considered the most contributed feature toward GWL prediction.

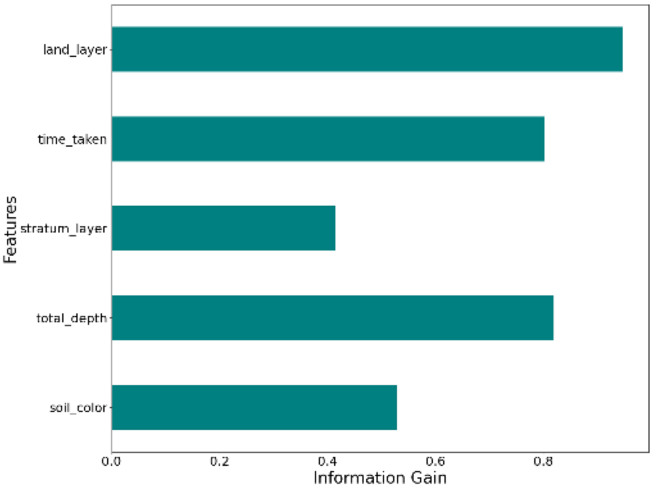


FIGURE 24. Features impact analysis using IG.

IX. CONCLUSION

The importance of groundwater level has received high significance due to variation in hydrogeological properties. The proposed ensemble prediction model was presented to develop an integrated prediction model based on boosting and bagging models using boreholes-log data to predict GWL for sustainable water resource planning and management. The proposed research study consists of two core modules; data and predictive analytics modules. The data analytics module aimed to process and investigate boreholes data to discover hidden hydrogeological characteristics to improve the efficiency of future boreholes. Therefore, different data analysis techniques were employed to analyze boreholes data, such as statistical and time-series analyses of borehole data based on soil colors, land layers, and stratum layers, to name a few. Differencing and correlation mechanisms were also utilized to find a difference between consecutive boreholes depths and analyze the linear relationship between boreholes depths. Furthermore, different hydrogeological and time interval features were extracted from the prepared boreholes log data. Secondly, the predictive analytics module aimed to develop an ensemble prediction model based on the integration of multiple boosting and bagging models using extracted hydrogeological and temporal features to predict GWL. The ultimate goal of the proposed ensemble prediction model was to predict GWL in order to facilitate drilling management for sustainable water resource management. Furthermore, prediction errors of the implemented models were evaluated using different error metrics. The MAE, MSE, RMSE, and NRMSE values of the proposed E-GWLP model are 0.340, 0.564, 0.751, and 0.018, respectively, which indicates that our E-GWLP model accurately predicted GWL compared to conventional ensemble and baseline regression models. The prediction error of the proposed ensemble model in terms of MAPE for unseen samples is 12.658, which signifies that E-GWLP model performed quite well compared to the baseline models. In contrast, MAPE of KNN-RF and XGB-RF is 34.048, and 28.444 respectively, which indicates that traditional hybrids models produced a relatively high prediction error compared to our proposed model. Furthermore, evaluation results of the E-GWLP model were compared with six conventional ML models, such as ANN, SVR, DT, LR, L1, and L2. The analysis of the traditional regression models shows that the LR model performed poorly compared to other baseline models. The experimental results revealed that the proposed E-GWLP model accurately predicts GWL and outperformed conventional regression models. The experimental results will be used for the planning and management of sustainable water resources. Moreover, It will also be used to improve reservoir engineering and the efficiency of future boreholes.

CONFLICT OF INTERESTS

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

REFERENCES

[1] M. Ehteram, V. P. Singh, A. Ferdowsi, S. F. Mousavi, S. Farzin, H. Karami, N. S. Mohd, H. A. Afan, S. H. Lai, O. Kisi et al., “An improved model based on the support vector machine and cuckoo algorithm for simulating reference evapotranspiration,” PloS one, vol. 14, no. 5, p. e0217499, 2019.

[2] 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.

[3] E. Merem, Y. Twumasi, J. Wesley, M. Alsarari, S. Fageir, M. Crisler, C. Romorno, D. Olagbegi, A. Hines, G. Ochai et al., "Assessing water resource issues in the us pacific north west region," in Proceedings of Mississippi Political Science Conference (MPCC). Jackson: Mississippi, February2018, 2018.

[4] I. Bremere, M. Kennedy, A. Stikker, and J. Schippers, "How water scarcity will effect the growth in the desalination market in the coming 25 years," Desalination, vol. 138, no. 1-3, pp. 7–15, 2001.

[5] K. E. Kemper, "Groundwater—from development to management," Hydrogeology Journal, vol. 12, no. 1, pp. 3–5, 2004.

[6] M. Ehteram, H. A. Afan, M. Dianatikhah, A. N. Ahmed, C. Ming Fai, M. S. Hossain, M. F. Allawi, and A. Elshafie, "Assessing the predictability of an improved anfis model for monthly streamflow using lagged climate indices as predictors," Water, vol. 11, no. 6, p. 1130, 2019.

[7] B. Hölting and W. G. Coldewey, "Groundwater exploitation," in Hydrogeology. Springer, 2019, pp. 203–230.

[8] Y. Guo, S. Dong, Y. Hao, Z. Liu, T.-C. J. Yeh, W. Wang, Y. Gao, P. Li, and M. Zhang, "Risk assessments of water inrush from coal seam floor during deep mining using a data fusion approach based on grey system theory," Complexity, vol. 2020, 2020.

[9] H. Niroumand, M. Zain, and M. Jamil, "Statistical methods for comparison of data sets of construction methods and building evaluation," Procedia-Social and Behavioral Sciences, vol. 89, pp. 218–221, 2013.

[10] C. Hegde, H. Daigle, H. Millwater, and K. Gray, "Analysis of rate of penetration (rop) prediction in drilling using physics-based and data-driven models," Journal of Petroleum Science and Engineering, vol. 159, pp. 295–306, 2017.

[11] E. Zaveri, D. S. Grogan, K. Fisher-Vanden, S. Frolking, R. B. Lammers, D. H. Wrenn, A. Prusevich, and R. E. Nicholas, "Invisible water, visible impact: groundwater use and indian agriculture under climate change," Environmental Research Letters, vol. 11, no. 8, p. 084005, 2016.

[12] M. E. Hossain, A. Al-Majed, A. R. Adebayo, A. S. Apaleke, and S. M. Rahman, "A critical review of drilling waste management towards sustainable solutions." Environmental Engineering & Management Journal (EEMJ), vol. 16, no. 7, 2017.

[13] G. Thonhauser, "Using real-time data for automated drilling performance analysis," European Oil and Gas Magazine, vol. 4, p. 170ff, 2004.

[14] H. Musbah, M. El-Hawary, and H. Aly, "Identifying seasonality in time series by applying fast fourier transform," in 2019 IEEE Electrical Power and Energy Conference (EPEC). IEEE, 2019, pp. 1–4.

[15] K. Du, Y. Zhao, and J. Lei, "The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series," Journal of Hydrology, vol. 552, pp. 44–51, 2017.

[16] R. C. Brasileiro, V. L. Souza, and A. L. Oliveira, "Automatic trading method based on piecewise aggregate approximation and multi-swarm of improved self-adaptive particle swarm optimization with validation," Decision Support Systems, vol. 104, pp. 79–91, 2017.

[17] S. Liu, H. Ji, and M. C. Wang, "Nonpooling convolutional neural network forecasting for seasonal time series with trends," IEEE transactions on neural networks and learning systems, vol. 31, no. 8, pp. 2879–2888, 2019.

[18] A. M. Alsalama, J. P. Canlas, S. H. Gharbi et al., "An integrated system for drilling real time data analytics," in SPE Intelligent Energy International Conference and Exhibition. Society of Petroleum Engineers, 2016.

[19] B. Esmael, A. Arnaout, R. K. Fruhwirth, and G. Thonhauser, "A statistical feature-based approach for operations recognition in drilling time series," International Journal of Computer Information Systems and Industrial Management Applications, vol. 5, pp. 454–61, 2015.

[20] 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.

[21] N. Iqbal, R. Ahmad, F. Jamil, and D.-H. Kim, "Hybrid features prediction model of movie quality using multi-machine learning techniques for effective business resource planning," Journal of Intelligent & Fuzzy Systems, no. Preprint, pp. 1–22.

[22] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, "Wr-svm model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification," Applied Sciences, vol. 11, no. 10, p. 4657, 2021.

[23] 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.

[24] N. Ahmad, L. Han, K. Iqbal, R. Ahmad, M. A. Abid, and N. Iqbal, "Sarm: salah activities recognition model based on smartphone," Electronics, vol. 8, no. 8, p. 881, 2019.

[25] 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.

[26] 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.

[27] H. Karami, S. F. Mousavi, S. Farzin, M. Ehteram, V. P. Singh, and O. Kisi, "Improved krill algorithm for reservoir operation," Water Resources Management, vol. 32, no. 10, pp. 3353–3372, 2018.

[28] M. Ehteram, H. Karami, S. F. Mousavi, S. Farzin, A. B. Celeste, and A.-E. Shafie, "Reservoir operation by a new evolutionary algorithm: Kidney algorithm," Water resources management, vol. 32, no. 14, pp. 4681–4706, 2018.

[29] S. Maroufpoor, A. Fakheri-Fard, and J. Shiri, "Study of the spatial distribution of groundwater quality using soft computing and geostatistical models," ISH Journal of Hydraulic Engineering, vol. 25, no. 2, pp. 232–238, 2019.

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

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

[32] M. Zare and M. Koch, "Groundwater level fluctuations simulation and prediction by anfis-and hybrid wavelet-anfis/fuzzy c-means (fcm) clustering models: Application to the miandarband plain," Journal of Hydro-environment Research, vol. 18, pp. 63–76, 2018.

[33] Y. Tang, C. Zang, Y. Wei, and M. Jiang, "Data-driven modeling of groundwater level with least-square support vector machine and spatial–temporal analysis," Geotechnical and Geological Engineering, vol. 37, no. 3, pp. 1661–1670, 2019.

[34] J. Adamowski and H. F. Chan, "A wavelet neural network conjunction model for groundwater level forecasting," Journal of Hydrology, vol. 407, no. 1-4, pp. 28–40, 2011.

[35] R. Taormina, K.-w. Chau, and R. Sethi, "Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the venice lagoon," Engineering Applications of Artificial Intelligence, vol. 25, no. 8, pp. 1670–1676, 2012.

[36] 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.

[37] H. Yoon, Y. Hyun, K. Ha, K.-K. Lee, and G.-B. Kim, "A method to improve the stability and accuracy of ann-and svm-based time series models for long-term groundwater level predictions," Computers & geosciences, vol. 90, pp. 144–155, 2016.

[38] 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.

[39] O. H. Kombo, S. Kumaran, Y. H. Sheikh, A. Bovim, and K. Jayavel, "Long-term groundwater level prediction model based on hybrid knn-rf technique," Hydrology, vol. 7, no. 3, p. 59, 2020.

[40] X. Wang, T. Liu, X. Zheng, H. Peng, J. Xin, and B. Zhang, "Short-term prediction of groundwater level using improved random forest regression with a combination of random features," Applied Water Science, vol. 8, no. 5, pp. 1–12, 2018.

[41] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, "Gis-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in iran," Environmental monitoring and assessment, vol. 188, no. 1, pp. 1–27, 2016.

[42] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," Water, vol. 11, no. 5, p. 910, 2019.

[43] V. M. Herrera, T. M. Khoshgoftaar, F. Villanustre, and B. Furht, "Random forest implementation and optimization for big data analytics on lexisnexis's high performance computing cluster platform," Journal of Big Data, vol. 6, no. 1, pp. 1–36, 2019.

[44] M. Zabihi, H. R. Pourghasemi, Z. S. Pourtaghi, and M. Behzadfar, "Gis-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in iran," Environmental Earth Sciences, vol. 75, no. 8, p. 665, 2016.

[45] P. Baudron, F. Alonso-Sarría, J. L. García-Aróstegui, F. Cánovas-García, D. Martínez-Vicente, and J. Moreno-Brotóns, "Identifying the origin of

groundwater samples in a multi-layer aquifer system with random forest classification," Journal of Hydrology, vol. 499, pp. 303–315, 2013.

[46] B. Li, G. Yang, R. Wan, X. Dai, and Y. Zhang, "Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the poyang lake in china," Hydrology Research, vol. 47, no. S1, pp. 69–83, 2016.

[47] M. Sakizadeh, M. M. Mohamed, and H. Klammler, "Trend analysis and spatial prediction of groundwater levels using time series forecasting and a novel spatio-temporal method," Water Resources Management, vol. 33, no. 4, pp. 1425–1437, 2019.

[48] E. H. de Moraes Takafuji, M. M. da Rocha, and R. L. Manzione, "Groundwater level prediction/forecasting and assessment of uncertainty using sgs and arima models: A case study in the bauru aquifer system (brazil)," Natural Resources Research, vol. 28, no. 2, pp. 487–503, 2019.

[49] B. Dhekale, P. Sahu, K. Vishwajith, and L. Narsimahaiah, "Structural time series analysis towards modeling and forecasting of groundwater fluctuations in murshidabad district of west bengal," Ecosystem, vol. 5, pp. 117–126, 2015.

[50] Z. Şen, "Innovative trend significance test and applications," Theoretical and applied climatology, vol. 127, no. 3-4, pp. 939–947, 2017.

[51] I. Minea, D. Boicu, and O.-E. Chelariu, "Detection of groundwater levels trends using innovative trend analysis method in temperate climatic conditions," Water, vol. 12, no. 8, p. 2129, 2020.

[52] 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, 2021.

[53] 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.

[54] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337). IEEE, 1999, pp. 126–133.

[55] J. K. Kruschke, "Bayesian estimation supersedes the t test." Journal of Experimental Psychology: General, vol. 142, no. 2, p. 573, 2013.

[56] S. Hido, T. Idé, H. Kashima, H. Kubo, and H. Matsuzawa, "Unsupervised change analysis using supervised learning," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2008, pp. 148–159.

[57] T. L. Saaty and M. Sagir, "An essay on rank preservation and reversal," Mathematical and Computer Modelling, vol. 49, no. 5-6, pp. 1230–1243, 2009.

[58] X. Shao and X. Zhang, "Testing for change points in time series," Journal of the American Statistical Association, vol. 105, no. 491, pp. 1228–1240, 2010.

[59] C. Croux and C. Dehon, "Influence functions of the spearman and kendall correlation measures," Statistical methods & applications, vol. 19, no. 4, pp. 497–515, 2010.

[60] S. Hörmann, P. Kokoszka, G. Nisol et al., "Testing for periodicity in functional time series," Annals of statistics, vol. 46, no. 6A, pp. 2960–2984, 2018.

[61] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, 2018.

[62] B. P. BV and M. Dakshayini, "Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data," Procedia computer science, vol. 132, pp. 198–207, 2018.

NAEEM IQBAL is currently pursuing Ph.D. in the Department of Computer Engineering at Jeju National University, Republic of Korea. He received his M.S in Computer Science from COMSATS University Islamabad, Attock Campus, Punjab, Pakistan in 2019. He did his B.S in Computer Science from the COMSATS University Islamabad, Attock Campus, Pakistan. 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.

ANAM-NAWAZ KHAN is currently pursuing Ph.D. at the Department of Computer Engineering Jeju National University, Republic of Korea. She received the M.S degree in Computer Science from COMSATS University Islamabad, Attock Campus, Pakistan in 2019. She did B.S in Computer Science from the COMSATS University Islamabad, Attock Campus in 2016. Her research work mainly focused on machine learning applications in smart environments, analysis of prediction and optimization algorithms, big data and IoT-based Applications.

ATIF RIZWAN is currently pursuing Ph.D. in the Department of Computer Engineering at Jeju National University, 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.

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, Republic of Korea, in 2015. Since 2016, he has been with COMSATS University Islamabad, Attock Campus, Pakistan, where he is currently an Assistant Professor with the Department of Computer Science. His 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 a principal researcher at Electronics and Telecommunications Research Institute (ETRI), working on development of edge computing system for solving urban problems. He received his BS degree in electronics engineering from Han-yang University, Republic of Korea, in 1992, his MS degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1994, and his Ph.D in computer science electrical engineering from KAIST in 2000. He is currently interested in edge computing, wireless sensor network, and data analysis based on artificial intelligence.

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

DOHYEUN KIM received the B.S. degree in electronics engineering from Kyungpook National University, Republic of Korea, in 1988, and the M.S.and Ph.D. degrees in information telecommunication from Kyungpook National University, Republic of 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, Republic of Korea, where he is currently a Professor with the Department of Computer Engineering. From 2008 to 2009, he was a Visiting Researcher with the Queensland University of Technology, Australia. His research interests include sensor networks, M2M/IOT, energy optimization and prediction, intelligent service, and mobile computing.

• • •