Predictive Modeling of Groundwater Recharge under Climate Change Scenarios in the Northern Area of Saudi Arabia

Rabie A. Ramadan
Computer Engineering Department, College of Computer Science and Engineering, Hail University, Saudi Arabia | Computer Engineering Department, Faculty of Engineering, Cairo University, Egypt
rabieramadan@gmail.com (corresponding author)

Sahbi Boubaker
Department of Computer and Network Engineering, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
sboubaker@uj.edu.sa

Received: 6 February 2024 | Revised: 22 February 2024 | Accepted: 1 March 2024

ABSTRACT

Water scarcity is considered a major problem in dry regions, such as the northern areas of Saudi Arabia and especially the city of Hail. Water resources in this region come mainly from groundwater aquifers, which are currently suffering from high demand and severe climatic conditions. Forecasting water consumption as accurately as possible may contribute to a high level of sustainability of water resources. This study investigated different Machine Learning (ML) algorithms, namely Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), and Gradient Boosting (GB), to efficiently predict water consumption in such areas. These models were evaluated using a set of performance measures, including Mean Squared Error (MSE), R-squared (R^2), Mean Absolute Error (MAE), Explained Variance Score (EVS), Mean Absolute Percentage Error (MAPE), and Median Absolute Error (MedAE). Two datasets, water consumption and weather data, were collected from different sources to examine the performance of the ML algorithms. The novelty of this study lies in the integration of both weather and water consumption data. After examining the most effective features, the two datasets were merged and the proposed algorithms were applied. The RF algorithm outperformed the other models, indicating its robustness in capturing water usage behavior in dry areas such as Hail City. The results of this study can be used by local authorities in decision-making, water consumption analysis, new project construction, and consumer behavior regarding water usage habits in the region.

Keywords-groundwater; recharge; Hail city; machine learning; prediction

I. INTRODUCTION

Today, many concerns are raised about environmental conditions and their effects on the global population. This raises the issue of water management as a natural resource. Water in countries such as Saudi Arabia, especially in northern areas such as Hail City, is a precious resource [1] and management and prediction of water consumption are challenging problems. This study uses Machine Learning (ML) capabilities in water management in the northern areas of Saudi Arabia, focusing on predictive analytics and using available datasets of water consumption and weather data [2-3]. Instead of traditional forecasting methods, which depend on historical data, this study aims to evaluate and compare four ML models: Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), and Gradient Boosting (GB) in the context of water consumption prediction. These models were selected because of their proven capabilities in similar forecasting frameworks when applied to complex multidimensional data.

Data collection is the first step of a successful prediction. In this regard, the two datasets used were collected from different sources [2-3]. This study uses a water consumption dataset, which contains records of monthly water consumption, and comprehensive hourly weather observations. Modeling the relationship between weather data and water consumption is crucial for managing water resources under climate change conditions, as water demand may increase following an increase in temperature [4]. Additionally, other weather data, such as wind speed and visibility, greatly affect water consumption.

This study aims to make a significant contribution to the prediction of water consumption in the northern area of Saudi Arabia, which has a special climate with a scarcity of rain and groundwater. Four methods were evaluated with the Mean Square Error (MSE), R square (R^2), Mean Absolute Error (MAE), Explained Variance Score (EVS), Mean Absolute Percentage Error (MAPE), and Median Absolute Error (MedAE). These measures were selected to reflect the accuracy and reliability of the proposed algorithms in forecasting groundwater consumption in northern Saudi Arabia. For example, MSE reflects the sensitivity of large errors, which critically assesses the accuracy of the prediction, whereas R^2 and EVS measure the proportion of variance in water consumption that the models could explain, indicating the fitting of the models used to the datasets and their ability to capture the underlying patterns in water usage. Therefore, they perfectly reflect the study's objective of developing a model that accurately reflects consumption dynamics. MAE and MedAE indicate the average magnitude and central tendency of these errors, respectively. These measures are less sensitive to outliers than MSE. Therefore, a more robust view of model performance in different scenarios is very important in water resource management. Finally, MAPE is a relative error metric that was chosen to reflect the perspective on the error size relative to actual consumption values and is particularly important to policymakers because it gives an intuitive understanding of model accuracy in percentage terms.

The results could benefit the local authorities of Saudi Arabia in managing water consumption in these areas and extend the results to many similar locations. However, in leveraging predictive analytics for water management, ethical integrity requires transparency, privacy protection, and equity, ensuring that the models do not perpetuate biases. Crucially, engaging communities in decision-making processes fosters trust and ensures that water management practices are inclusive and fair, respect diverse needs, and promote equitable access to water resources.

II. RELATED WORKS

The prediction of water consumption and Groundwater Level (GWL) plays a crucial role in water resource management. Therefore, researchers and water systems practitioners have been very interested in this topic for the last few decades. Traditional forecasting techniques have shown several limitations, although the potential they exhibited when applied to water time series. Some of these methods have linear relationships between variables. However, water consumption and GWLs often exhibit non-linear behavior patterns. Additionally, many techniques frequently require the time series data to be stationary, meaning the statistical properties of the series (mean, variance, autocorrelation, etc.) are constant over time. Additionally, they may focus on historical data patterns and not effectively incorporate external factors or variables. Moreover, they face difficulty in handling large and complex datasets. On the other hand, machine learning techniques consider the variability in the dataset and can handle large datasets.

Water consumption greatly affects water resources, including GWLs, since climatic, topographic, and hydrological factors interact, making simulation and prediction challenging. In [5], a comprehensive overview of soft computing techniques, including ML, was provided to predict GWLs while considering the techniques used, the study location, the dataset, and the main performance metrics. This study showed that the prediction of GWLs is case-sensitive. In [6], a real-time framework was presented to predict short-term water consumption, based on a back-propagation ML algorithm and Long-Short-Term Memory (LSTM). The results obtained were promising, as they were found to have an accuracy of a few liters. Combining human mobility, historical water consumption, and applying ML algorithms such as SVM and tree-based ensemble methods yielded accurate forecasts for a case study in Wroclaw, Poland [7]. Compared to classical techniques such as ARIMA, ML techniques have improved accuracy. In [8], the water irrigation demand was estimated based on an RF algorithm. The results provided a projected spatial resolution of 6 km, considering evapotranspiration and meteorological data.

GWL prediction, as among the drivers of water consumption, has been extensively considered using ML algorithms [9-17]. GWL is in mutual interaction with water demand/consumption, mainly in dry regions with extensive agricultural activities such as Hail City, Saudi Arabia. GWL is affected by several factors, including climatic conditions [9] and regional characteristics [10]. The groundwater depth usually exhibits fluctuations due to climate, water extraction/demand, surface water flows, and rainfall. The relationship between GWL and these factors is known to be highly nonlinear. The classic techniques usually fail to capture those complex relationships. In [11], ML methods were used to model these relationships, and the tested Deep Learning (DL) models provided good results, although they suffered from relatively limited datasets. Authors in [12] highlighted that the studied aquifer was in a critical situation and recommended urgent management actions. A new trending method to predict GWL consists of using ensemble DL methods [13], as the results have shown their superiority over single methods. Combined methods have also been investigated while using different techniques, sometimes from different categories, such as ML with wavelet transforms [14]. ML and DL have also been investigated in the domain of water prediction [15-17]. However, case studies from arid regions with scarce water resources and difficult climatic conditions, such as in the north of Saudi Arabia, were not adequately investigated.

This study fills this research gap by investigating water consumption in the northern area of Saudi Arabia using different algorithms, including RF, GB, SVM, and LR. Similar frameworks to these methods can be found in [18-24]. The novel contributions of this study are that it is carried out for a special area with special characteristics, it also merges two datasets for efficient water consumption prediction, including water and weather data, and finally applies different ML algorithms to the merged datasets.

III. PROPOSED SOLUTION APPROACH

Figure 1 shows the proposed approach, which is broken down into five key steps:

- Data collection: At this stage, data are collected from different sources [2-3]. The collected water data include historical observations of water consumption and historical weather data. The quality of the data collected certainly affects the accuracy of the prediction process.

- Data Processing: It is a critical step that cleans the data and transforms them into a suitable format for analysis. Preprocessing operations include finding missing values, normalization, and selecting relative features. Preprocessing also involves recognizing seasonal variations and timestamp conversion.
- Data Split: In this phase, the data are divided into training and testing sets. The training data are used to train the investigated models, and the test data are used to test the accuracy of the prediction.
- Apply Models: ML models are applied to the data. Each model requires specific data handling. This step is critical to evaluate the performance of the proposed models.
- Evaluation of the models: In the last phase, the results are evaluated. Different evaluation methods were chosen, such as MSE, R^2, MAE, EVS, MAPE, and MedAE.

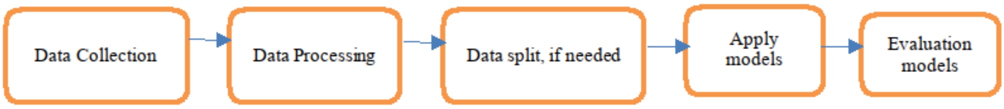

Fig. 1. Proposed approach.

This study used the following ML models.

A. Random Forest (RF) Regressor

RF is an ensemble learning technique mostly used for classification and regression tasks [18]. It constructs multiple trees during the training phase, based on the type of data, and gives a chance to avoid overfitting, which is common for individual decision trees. It does this by averaging over a multitude of decision trees. The algorithm is also suitable for complex and nonlinear data. It also provides valuable information about the feature's importance, leading to the selection of important and effective features. In addition, it can handle various data types, which makes it suitable for ML applications.

B. Gradient Boosting (GB) Regressor

GB is known for its computational efficiency and performance, proven in large- and complex datasets. It involves regularization (L1 and L2) that helps avoid overfitting. An important feature of GB is its efficient handling of missing data, which makes it robust for modeling without extensive preprocessing. Moreover, it has tree pruning, which limits tree growth during the learning process [19].

C. Linear Regression (LR)

LR is an ML technique that is used to model the relationship between two variables by fitting a linear equation to the observed data. Some variables are considered explanatory, and another variable is considered the dependent variable. In this study, the water demand is considered the dependent variable, and the explanatory variables are the climatic conditions [25].

D. Support Vector Machines (SVMs)

SVMs consist of a group of supervised learning methods that are usually used in classification and regression problems. SVMs are extended to nonlinear models of the generalized algorithm developed for linear regressions [20].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The RF was designed with 100 estimators and a fixed random state for reproducibility. The RF model, as shown in Figure 2, showed reasonable results, with an R^2 of approximately 0.694. This indicates that the model was able to explain 69.4% of the variance in water consumption. At the same time, MSE was 0.549 and MAE was 0.529, which are relatively low. In addition, a substantial portion of the variance in water consumption was shown through EVS, which was 0.694. The MAPE of 3.05% indicates a lower prediction error, while the MSLE of 0.0016 and MedAE of 0.371 again confirm its accuracy.

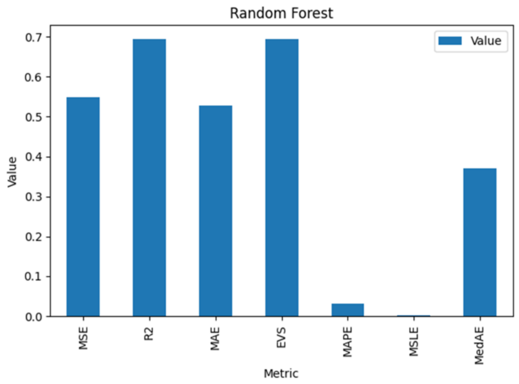

Fig. 2. RF evaluation metrics.

Figure 3 shows the results of the LR, indicating that it is a moderate fit to predict water consumption. LR had R^2, MSE, MAE, EVS, MAPE, and MAE of 0.470, 0.951, 0.846, 0.470, 4.89%, and 0.808, respectively. These results indicate that the model was able to explain nearly 47% of the variance in water consumption with a moderate level of accuracy. The results assume that the model accounts for a substantial portion of the variance in water consumption and offers a robust assessment of central prediction errors. The SVM results, shown in Figure 4, indicate weak performance, as R^2 was approximately 0.0039, indicating that the model captured only 0.39% of the variance in water consumption. This evaluation was also confirmed by MAE, EVS, MAPE, MSLE, and MedAE. The GB was configured similarly to the RF with 100 estimators.

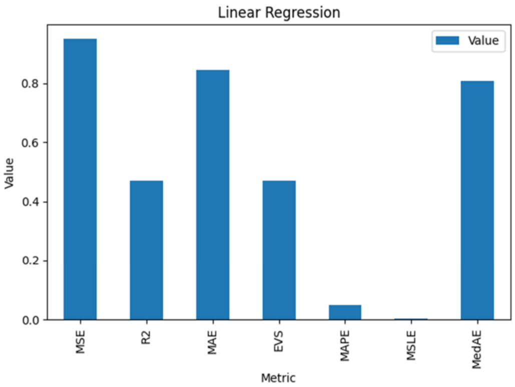

Fig. 3. LR evaluation metrics.

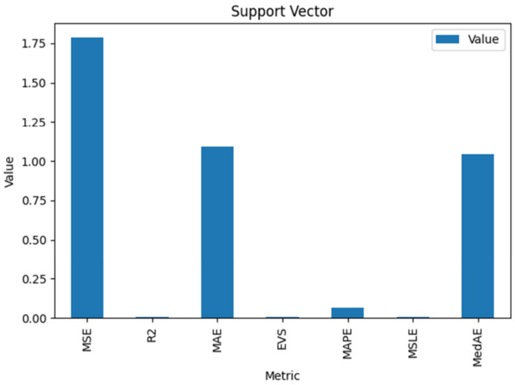

Fig. 4. SVM evaluation metrics.

Figure 5 shows the GB loss curve, plotting MSE against the number of estimators for training and validation data. The figure suggests effective learning with increasing estimators. As the number of estimators continues to increase, the curve plateau indicates diminishing returns on error reduction, and the model is generalizing well and not overfitting. Figure 6 presents the comparison between the actual and the predicted values from a GB model. The red (predicted) and blue (actual) points show the overall accuracy of the model. Figure 7 shows the performance of the GB model, showing an R^2 score of 0.596, indicating that it can account for approximately 59.63% of the variance in water consumption. Similarly, MSE equals 0.724, implying that the model's predictions have a relatively moderate level of accuracy. Also, MAE of 0.711 and EVS of 0.596 indicate that, on average, the model's predictions deviate by this amount from the actual values. Other measures confirm the given results.

Table I summarizes the evaluation results of the considered algorithms. RF outperformed the other models in almost all metrics. GB follows closely behind, showing robust performance. LR and SVM lagged, with LR performing moderately, and SVM showed the worst results. Although RF shows the best results among the other models, its accuracy is still not up to the expected results. These results reflect the complexity of the prediction based on two different datasets that were not collected for water consumption. To the best of our knowledge, no previous study has combined weather and water data, especially in the northern areas of Saudi Arabia. However, this study is a step forward in the relationship between weather data and water use and consumption in different areas.

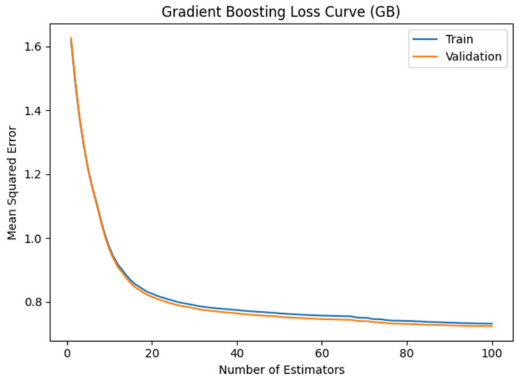

Fig. 5. GB loss curve.

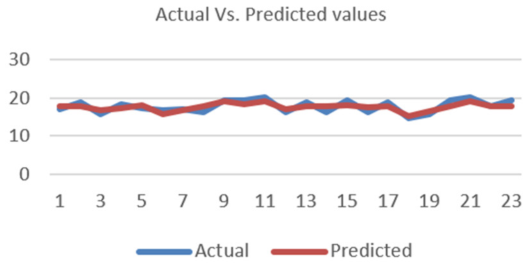

Fig. 6. GB actual vs predicted values.

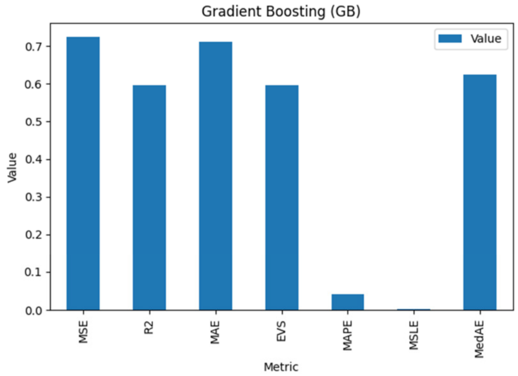

Fig. 7. GB performance metrics.

TABLE I. PERFORMANCE COMPARISON

Algorithm	**MSE (lt^2)**	**R^2(%)**	**MAE (lt)**	**EVS (unitless)**	**MAPE(%)**	**MedAE (lt)**
SVM	1.7864	0.0039	1.0929	0.0053	0.0634	1.0411
RF	0.5489	0.6939	0.5288	0.6939	0.0305	0.3705
LR	0.9506	0.4699	0.8461	0.4700	0.0489	0.8081
GB	0.7240	0.5962	0.7111	0.5963	0.0411	0.6239

V. CHALLENGES AND LIMITATIONS

Studying water consumption in areas such as northern regions of Saudi Arabia faces challenges such as data scarcity and quality, reliance on non-renewable groundwater, climate change, socioeconomic growth, technological and infrastructural limitations, and environmental concerns. This area has limited historical quality datasets, which hinder accurate modeling and forecasting of water usage. This study had to collect data from various sources to obtain enough for the investigation. Also, the combination of underground water and weather data demands careful resource management. Additionally, government policies and the need to balance ecological sustainability with human water use present complex challenges. The complexity of predicting water consumption is increased by the interplay of various external factors, making accurate forecasting challenging.

VI. IMPLICATIONS OF THE FINDINGS FOR WATER RESOURCE MANAGEMENT POLICIES

The results of this study can have a great impact on water resource management policies in northern Saudi Arabia, highlighting the critical role of predictive analytics in the formation of sustainable strategies. Using predictive analytics leads to a better understanding of patterns of water consumption and prediction of future demands. Therefore, policymakers could have results to make informed decisions. The results of this research identify the key factors driving water consumption in the region. Policies can target these factors to reduce water consumption and usage. Adaptive analytics could optimize the allocation of water resources. Therefore, policymakers can ensure that water distribution is managed more efficiently, decreasing wastage and ensuring water is available where and when it is most needed. Intelligent irrigation systems for agriculture could be developed, ensuring that water is used precisely and efficiently to meet crop needs without excess. Finally, the results of this study highlight the importance of integrating predictive analytics into long-term water resource planning and infrastructure development. Therefore, policymakers can better plan for the infrastructure of water storage and distribution.

VII. CONCLUSION

The results of this study underscore the potential of ML techniques, especially RF, in predicting water consumption in areas such as northern Saudi Arabia and Hail City. RF showed superior performance with the lowest MSE and the highest R^2 and EVS, indicating its ability to handle non-linear and interesting water usage patterns. However, SVM was unable to integrate and perform well in the given water and weather datasets. This study is one of the first that integrates weather data into water usage data in such areas, where weather and water consumption usage and sources are exceptional cases. This study could be extended to incorporate the interaction between climate conditions and agricultural activities or between underground water consumption and consumer behavior. Future research will focus on the use of different algorithms, including DL algorithms, to refine the prediction. Also, real-time analytics is expected to have more value than offline prediction.

ACKNOWLEDGMENT

This research was funded by the Ali Aljomaiya Scientific Research Chair for Sustainable Development at the University of Hail, Saudi Arabia, through Project Number SCR-22 045.

REFERENCES

[1] A. I. Almulhim and I. R. Abubakar, "Developing a sustainable water conservation strategy for Saudi Arabian cities," *Groundwater for Sustainable Development*, vol. 23, Nov. 2023, Art. no. 101040, https://doi.org/10.1016/j.gsd.2023.101040.

[2] "Saudi Arabia Hourly Climate Integrated Surface Data." https://datasource.kapsarc.org/explore/dataset/saudi-hourly-weather-data/information.

[3] "Per Capita Water Consumption In Saudi Regions." https://datasource.kapsarc.org/explore/dataset/per_capita_average_water_use_in_saudi_regions/information.

[4] D. Fiorillo, Z. Kapelan, M. Xenochristou, F. De Paola, and M. Giugni, "Assessing the Impact of Climate Change on Future Water Demand using Weather Data," *Water Resources Management*, vol. 35, no. 5, pp. 1449–1462, Mar. 2021, https://doi.org/10.1007/s11269-021-02789-4.

[5] H. Tao *et al.*, "Groundwater level prediction using machine learning models: A comprehensive review," *Neurocomputing*, vol. 489, pp. 271–308, Jun. 2022, https://doi.org/10.1016/j.neucom.2022.03.014.

[6] A. Boudhaouia and P. Wira, "A Real-Time Data Analysis Platform for Short-Term Water Consumption Forecasting with Machine Learning," *Forecasting*, vol. 3, no. 4, pp. 682–694, Dec. 2021, https://doi.org/10.3390/forecast3040042.

[7] K. Smolak, B. Kasieczka, W. Fialkiewicz, W. Rohm, K. Siła-Nowicka, and K. Kopańczyk, "Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models," *Urban Water Journal*, vol. 17, no. 1, pp. 32–42, Jan. 2020, https://doi.org/10.1080/157306'2X.2020.1734947.

[8] S. Wei, T. Xu, G. Y. Niu, and R. Zeng, Estimating Irrigation Water Consumption Using Machine Learning and Remote Sensing Data in Kansas High Plains," *Remote Sensing*, vol. 14, no. 13, Jan. 2022, Art. no. 3004, https://doi.org/10.3390/rs14133004.

[9] F. B. Banadkooki *et al.*, "Enhancement of Groundwater-Level Prediction Using an Integrated Machine Learning Model Optimized by Whale Algorithm," *Natural Resources Research*, vol. 29, no. 5, pp. 3233–3252, Oct. 2020, https://doi.org/10.1007/s11053-020-09634-2.

[10] H. Cai, H. Shi, S. Liu, and V. Babovic, "Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States," *Journal of Hydrology: Regional Studies*, vol. 37, Oct. 2021, Art. no. 100930, https://doi.org/10.1016/j.ejrh.2021.100930.

[11] D. Kumar, T. Roshni, A. Singh, M. K. Jha, and P. Samui, "Predicting groundwater depth fluctuations using deep learning, extreme learning machine and Gaussian process: a comparative study," *Earth Science Informatics*, vol. 13, no. 4, pp. 1237–1250, Dec. 2020, https://doi.org/10.1007/s12145-020-00508-y.

[12] H. Kardan Moghaddam, S. Ghordoyee Milan, Z. Kayhomayoon, Z. Rahimzadeh kivi, and N. Arya Azar, "The prediction of aquifer groundwater level based on spatial clustering approach using machine learning," *Environmental Monitoring and Assessment*, vol. 193, no. 4, Mar. 2021, Art. no. 173, https://doi.org/10.1007/s10661-021-08961-y.

[13] H. A. Afan *et al.*, "Modeling the fluctuations of groundwater level by employing ensemble deep learning techniques," *Engineering Applications of Computational Fluid Mechanics*, vol. 15, no. 1, pp. 1420–1439, Jan. 2021, https://doi.org/10.1080/19942060.2021.1974093.

[14] A. T. M. S. Rahman, T. Hosono, J. M. Quilty, J. Das, and A. Basak, "Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms," *Advances in Water Resources*, vol. 141, Jul. 2020, Art. no. 103595, https://doi.org/10.1016/j.advwatres.2020.103595.

[15] M. Sapitang, W. M. Ridwan, A. N. Ahmed, C. M. Fai, and A. El-Shafie, "Groundwater level as an input to monthly predicting of water level

using various machine learning algorithms," *Earth Science Informatics*, vol. 14, no. 3, pp. 1269–1283, Sep. 2021, https://doi.org/10.1007/s12145-021-00654-x.

[16] W. Liu, H. Yu, L. Yang, Z. Yin, M. Zhu, and X. Wen, "Deep Learning-Based Predictive Framework for Groundwater Level Forecast in Arid Irrigated Areas," *Water*, vol. 13, no. 18, 2021, https://doi.org/10.3390/w13182558.

[17] W. Li, M. M. Finsa, K. B. Laskey, P. Houser, and R. Douglas-Bate, "Groundwater Level Prediction with Machine Learning to Support Sustainable Irrigation in Water Scarcity Regions," *Water*, vol. 15, no. 19, Jan. 2023, Art. no. 3473, https://doi.org/10.3390/w15193473.

[18] F. Mlawa, E. Mkoba, and N. Mduma, "A Machine Learning Model for detecting Covid-19 Misinformation in Swahili Language," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10856–10860, Jun. 2023, https://doi.org/10.48084/etasr.5636.

[19] M. Sipper and J. H. Moore, "AddGBoost: A gradient boosting-style algorithm based on strong learners," *Machine Learning with Applications*, vol. 7, Mar. 2022, Art. no. 100243, https://doi.org/10.1016/j.mlwa.2021.100243.

[20] K. Theofilatos, S. Likothanassis, and A. Karathanasopoulos, "Modeling and Trading the EUR/USD Exchange Rate Using Machine Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 2, no. 5, pp. 269–272, Oct. 2012, https://doi.org/10.48084/etasr.200.

[21] S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo, "Cost-sensitive probabilistic predictions for support vector machines," *European Journal of Operational Research*, vol. 314, no. 1, pp. 268–279, Apr. 2024, https://doi.org/10.1016/j.ejor.2023.09.027.

[22] R. G. Siqueira *et al.*, "Modelling and prediction of major soil chemical properties with Random Forest: Machine learning as tool to understand soil-environment relationships in Antarctica," *CATENA*, vol. 235, Feb. 2024, Art. no. 107677, https://doi.org/10.1016/j.catena.2023.107677.

[23] X. C. Nguyen *et al.*, "Estimating ammonium changes in pilot and full-scale constructed wetlands using kinetic model, linear regression, and machine learning," *Science of The Total Environment*, vol. 907, Jan. 2024, Art. no. 168142, https://doi.org/10.1016/j.scitotenv.2023.168142.

[24] S. Mondal, S. Ghosh, and A. Nag, "Brain stroke prediction model based on boosting and stacking ensemble approach," *International Journal of Information Technology*, vol. 16, no. 1, pp. 437–446, Jan. 2024, https://doi.org/10.1007/s41870-023-01418-0.