Forecasting Underground Water Levels: LSTM Based Model Outperforms GRU and Decision Tree Based Models

Md. Jafril Alam
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
jafrilalamshihab.kuetcse@gmail.com

Sujoy Kar
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
sujoykar2009@gmail.com

Sakib Zaman
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
sakibzaman169@gmail.com

Shamim Ahamed
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
shamim.pavel21@gmail.com

Kamrunnesa Samiya
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
samiya1707087@stud.kuet.ac.bd

***Abstract*—Underground water is too important in human life from various perspectives, but water depletion is a significant problem worldwide. The underground water level is decreasing daily and often increasing in some places. In order to save water, it is essential to observe underground water levels, so we proposed a methodology based on machine learning and deep learning to forecast underground water levels. This study used machine learning algorithms, including Random Forest and XGBoost Regressor, and deep learning algorithms like LSTM and GRU. We also looked into the effects of such algorithms on time series data. In terms of Mean absolute error (MAE) and Root Mean Squared Error (RMSE) score, the LSTM-based model outperformed the GRU-based model, Random Forest Regressor, and XGBoost Regressor. In our study, deep learning algorithm-based models outperformed traditional machine learning algorithms on time series data of water level. Specifically, LSTM outperformed other models in estimating groundwater levels.**

***Index Terms*—Underground water, Water level, Time series data, Machine Learning , Deep Learning , LSTM, GRU , Random Forest , XGBoost**

I. Introduction

Underground water is essential to humans in various aspects, including agriculture, supply of water, and industry [1]. However, the underground water level is falling daily, and underground water depletion is a worldwide issue [2]. Water waste and unrestricted water pumping deplete underground water. Additionally, the increasing underground water levels are sometimes caused by river flow and rainfall. As a result of depletion, the agricultural system and the production of industries can be impeded. If groundwater levels fall below a threshold level, there will be a shortage of drinking water. Therefore, monitoring the underground water level is crucial for saving water. Future water levels can be predicted using time series data of the water level of a specific period. Machine learning and deep learning-based models are used in data analysis and forecasting time series. However, the impact of different machine learning algorithms on time series forecasting varies.

II. Related works

Many studies on groundwater level predictions have already been conducted. Haiping Lin et al. [3] suggested a groundwater level forecasting system based on GRU. Stephanie R. Clark et al. [4] provided a method for comparing MLP, global MLP, SOM, LSTM, and DeepAR algorithms. S. Sahoo et al. [5] proposed a methodology that used ANN and regression models.

979-8-3503-1156-3/22/$31.00 © 2022 IEEE

III. The motive and goal of our work

The underground water levels are decreasing daily, causing drought and water scarcity. Therefore, it is critical to observe the cause of falling water levels and observe water levels to take appropriate steps to save underground water. Our first goal was to create models based on machine learning and deep learning to forecast underground water levels and select the best model. The second purpose was to examine whether deep learning-based or classical machine learning-based models performed better on time series data of underground water levels.

IV. Methodology

Figure-1 shows the proposed methodology of our research work. It is already described data processing and model evaluation metrics. The most concerning parts of the research were classical machine learning and deep learning-based architectures: LSTM, GRU, Random Forest Regressor, and XGBoost Regressor.

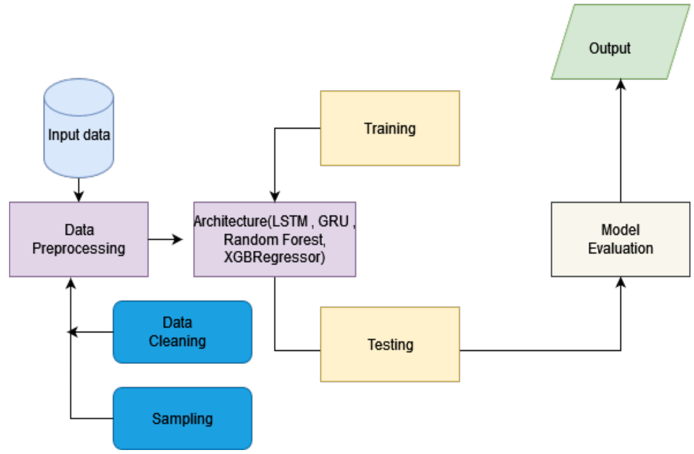

Fig. 1. Proposed methodology to forecast underground water levels

A. LSTM based model

Figure-2 depicts our proposed LSTM-based model, which includes the LSTM layer, Dense layer, drop-out, and tanh activation function. The Long short-term memory (LSTM) is one kind of recurrent neural network [6] that can handle sequence and time series data. An LSTM has both long-term and short-term memories and has three gates: an input gate, a forget gate, and an output gate. The cell's status is regulated by gates, with the input gate accepting input, the output gate producing output, and the forget gate erasing old data [7]. Firstly Data, previous hidden, and the internal state are used as input in LSTM. Then values of gates, current state, and current hidden state are calculated.

A dense layer is deeply connected with its preceding layer. The tanh activation function was used in this model. The tanh activation function is formulated as :

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

Here, f(x) denotes the outcome of the activation function, which denotes the predicted water level, and x is the input of the activation function.

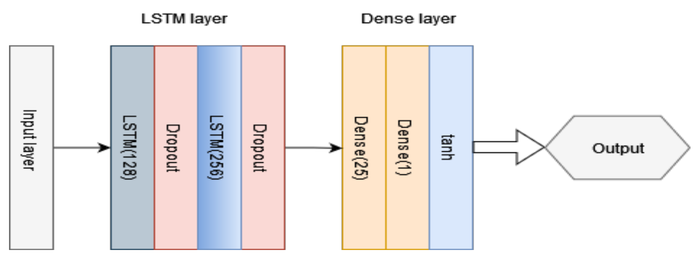

Fig. 2. Proposed LSTM architecture

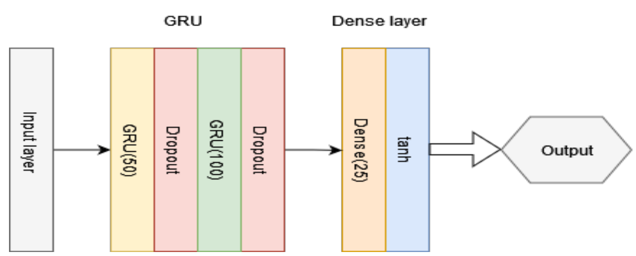

Fig. 3. Proposed GRU architecture

B. GRU based model

Figure-3 depicts our proposed GRU-based model, which contains the GRU layer, Dense layer, drop-out, and tanh activation function. The GRU is a form of recurrent neural network with two gates; whereas the LSTM has three gates, a GRU has two gates: the reset gate and the update gate [7] [8]. The reset gate is used to forget perception, whereas the update gate is used to remember based on perception. GRU's basic work procedures are as follows:

- Current data, previous hidden state are used as input
- The values of gates are calculated
- Current memory and hidden state are calculated

In this model tanh was used as activation function .

C. Random Forest

Breiman et al. [9] proposed the random forest algorithm, a tree-based classification algorithm currently utilized in time series data. [10] The random forest algorithm's basic operating procedures are as follows:

- Random Sampling of data
- Constructing tree and voting process for prediction

D. XGBoost Regressor

XGB, often known as XGBoost, is a scalable tree-based boosting technique for classification [11]. It can, however, be applied to time series data. It calculates residual and regularization parameters to produce output.

A comparison of these model are given below .

TABLE I
A COMPARISON AMONG THE MODELS

Algorithm	Basis	Suitable Data Size
LSTM	RNN	Large
GRU	RNN	Medium
RGBoost	Decision Tree	Medium
Random Forest	Decision Tree	Large

V. EXPERIMENT

A. Data

We used a time series dataset provided by the ACEA group from Kaggle to conduct this research. We updated this dataset slightly after gathering it for ease of use. The collection contains information about the some waterbody, which are aquifers. Those is located in Italy. The dataset has 49,233 observations in total. The dataset's key features include rainfall, temperature, river hydrometry, and drainage. Furthermore, the target column of the dataset is groundWater depth. Therefore, depending on factors affecting the subterranean water level, we needed to anticipate the water level for the next several days. The datasets used in this research is available in https://www.kaggle.com/competitions/acea-water-prediction/data .

B. Data Preprocessing and Analysis

Data preprocessing is critical in machine learning because untidy or unprepared data can lead to errors and incorrect results. So, firstly, we eliminated unnecessary and outdated features. Next, we dealt with missing values and performed a downsampling process on the dataset. Then we sorted the data in chronological order to maintain observation order. Finally, we visualized data and examined statistics as part of the data analysis.

C. Performance metric

Different types of metrics are utilized to evaluate the machine learning models. We measured the error of our model using the mean absolute error (MAE), and the root mean square error (RMSE) metrics.
MAE is formulated as :

$$MAE = \frac{1}{N}\sum_{j=1}^{N}(y_{pred}[j] - y_{act}[j]) \tag{2}$$

RMSE is formulated as :

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_{pred}[j] - y_{act}[j])^2} \tag{3}$$

where:

- $y_{pred}[j]$ is the predicted output
- $y_{act}[j]$ is the actual value
- N is total samples

D. Experimental Setup

In order to complete the experiment, we used a computer with a Windows operating system, an NVIDIA GEFORCE GPU, and 8 GB of RAM. We have used Kaggle and Colab as environments too, and most of the operations related to the experiment were executed in Colab.

E. Training the models

The models were trained after data preprocessing and model selection. Because LSTM and GRU are deep learning algorithms, the training strategies for all models were not the same. XGBoost and Random Forest, on the other hand, are classical machine learning-based algorithms. Although deep learning is a subset of machine learning, we classified algorithms into two groups to facilitate understanding. The following is a summary of models and critical parameters tuned during training:

TABLE II
SUMMARY OF TWO DEEP LEARNING BASED MODEL

Name	LSTM model	GRU model
No. of Dense layer	2	1
Activation Function	tanh	tanh
Dropout	0.2	0.2
Epoch	100	100
Folds	5	5

TABLE III
SUMMARY OF TWO CLASSICAL MACHINE LEARNING BASED MODEL

Name	XGBoost Regressor	Random Forest
Max Depth	4	5
Random state	0	42
No of Splits	3	3

VI. EXPERIMENTAL RESULTS

After the final executions of programs for our models, predicted results, MAE, and RMSE were recorded, and graphs were saved. MAE of the LSTM-based model, GRU-based model, Random Forest Regressor, and XGBRegressor were 0.144, 0.168, 1.493, and 1.982, respectively. RMSE of the LSTM-based model, GRU-based model, Random Forest Regressor, and XGBoost Regressor were 0.189, 0.203, 1.837, and 2.467, respectively. Table-IV shows the summary of the result.

TABLE IV
EXPERIMENTAL RESULTS FOR ALL ALGORITHMS BASED MODELS

Architecture	MAE	RMSE
LSTM	0.144	0.189
GRU	0.168	0.203
Random Forest	1.493	1.837
XGBRegressor	1.982	2.467

Figures-4 and 5 represent the graph of Groundwater depth vs. time for the LSTM-based and GRU-based models, respectively, where the blue line denotes the training set, the red line denotes predicted values, and the purple line denotes ground truth. In our research, LSTM based model performed best whereas XGBoost Regressor performed worst. It is clear that, for time series forecasting, deep learning-based models performed best over traditional machine learning algorithms. The dataset was large, so LSTM performed better than GRU because, for large data, LSTM works well.

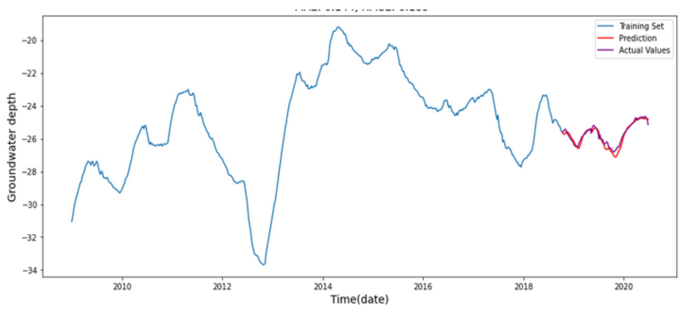

Fig. 4. Prediction graph of LSTM based model

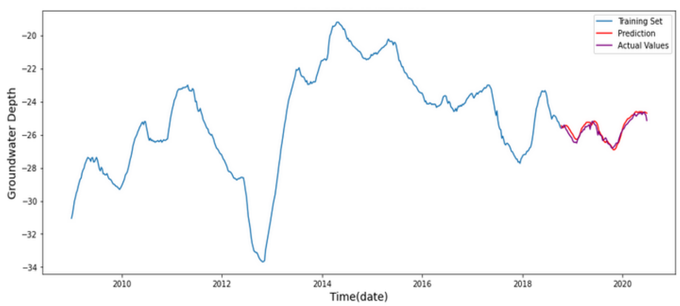

Fig. 5. Prediction graph of GRU based model

VII. Limitations and Future Works

Some limitations were discovered after completing this research work. Firstly, some vital features, such as pumped water, could be added. The second issue was certain dataset restrictions. The dataset was for a specific area, so we want to create a dataset of locations where the underground water level is quickly falling. Another thing to consider is developing an attention mechanism-based system to forecast time series data.

VIII. Conclusion

A system was designed and constructed successfully to forecast the underground water level. The system contains data processing, prediction, and model evaluation. As a result, the system can predict the future water level and produce predicted results and models error. The system employed various classical machine learning and deep learning-based algorithms. The LSTM outperformed the others in estimating underground water levels. It is also clear that deep learning-based models perform better than classical machine learning algorithms on time series data. Therefore, this system will help to observe water levels in many countries, and it will help to reduce the waste of water.

References

[1] Brands, Edwin, et al. "Groundwater." International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology (2016): 1-17.

[2] Konikow, Leonard F., and Eloise Kendy. "Groundwater depletion: A global problem." Hydrogeology Journal 13.1 (2005): 317-320.

[3] Lin, Haiping, et al. "Time series-based groundwater level forecasting using gated recurrent unit deep neural networks." Engineering Applications of Computational Fluid Mechanics 16.1 (2022): 1655-1672.

[4] Clark, Stephanie R., Dan Pagendam, and Louise Ryan. "Forecasting Multiple Groundwater Time Series with Local and Global Deep Learning Networks." International Journal of Environmental Research and Public Health 19.9 (2022): 5091.

[5] Sahoo, S., et al. "Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US." Water Resources Research 53.5 (2017): 3878-3895.

[6] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[7] Staudemeyer, Ralf C., and Eric Rothstein Morris. "Understanding LSTM–a tutorial into long short-term memory recurrent neural networks." arXiv preprint arXiv:1909.09586 (2019).

[8] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[9] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[10] Davis, Richard A., and Mikkel S. Nielsen. "Modeling of time series using random forests: Theoretical developments." Electronic Journal of Statistics 14.2 (2020): 3644-3671.

[11] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.