International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Performance and Robustness Analysis of Advanced Machine Learning Models for Predicting the Required Irrigation Water Amount

Hamed LAOUZ[1], Soheyb AYAD[1], Sadek Labib TERRISSA[1], Aicha Nabila BENHARKAT[2] and Samir MERDACI[3]

[1]*LINFI Laboratory, University Mohamed Khider, Biskra, Algeria*
[2]*Institut National des Sciences Appliquées de Lyon, Lyon, France*
[3]*Agronomy Faculty, University of El-oued, El-oued, Algeria*

Abstract: The agricultural sector plays a pivotal role in ensuring global food security, particularly in light of significant population growth. The demand for food is increasing substantially, while crop production may not sufficiently meet these rising needs. Water scarcity is one of the main problems that poses a significant challenge to the agriculture sector, exacerbated by inefficiencies in traditional irrigation methods. Addressing this issue requires accurate prediction of the precise water requirements of plants. In this paper, we introduce various machine learning and deep learning models designed to assess the water needs of greenhouse plants using daily changes in air environment and soil data. Results indicate that the Multi-Layer Perceptron (MLP) model consistently outperformed other models, demonstrating stability and efficacy across various data optimization phases. Additionally, Machine Learning (ML) and Long-Short Term Memory (LSTM) models displayed commendable performance in different data optimization scenarios. Robustness is used as a critical factor by analyzing the parameter sensitivity of each model. This analysis aids in comprehending the model's robustness before any model deployment. The results reveal the superior robustness of ML models compared to Deep Learning (DL) models. This robustness stems from the limited number of parameters utilized in ML models, enhancing their reliability in comparison to the proposed DL models.

Keywords: Precision irrigation, Water amount prediction, Data-based optimization, Hyper-parameters tuning, DL time series, Sensitivity analysis

1. INTRODUCTION

The global population reached 8 billion people in November 2022 and is expected to reach 9 billion by 2050 [1]. This has led to challenges, especially those related to food supply and freshwater resources, as water has become a critical resource that requires precise management, specifically in agriculture [2]. According to a report by the Food and Agriculture Organization (FAO) over the past century, global water use has increased at a rate that is more than twice as fast as population growth. In arid areas, population growth and economic development are putting unprecedented pressure on renewable but limited water supplies. Two-thirds of the world's population may be living in " stressful" circumstances by 2025, with 1.8 billion people predicted to reside in areas with "absolute" water shortage (Defined as less than $500m^3$ of water per person annually) and between 500 and $1000m^3$ of water per person annually [3].

The agricultural sector significantly contributes to water scarcity. According to an FAO report, 70% of the available freshwater resources are used for agriculture, and 60% of this water is wasted due to inefficient irrigation techniques [4]. These issues call for an innovative solution to rationalise water resource usage.

Numerous solutions have been developed to manage the amount of irrigation water used. Some utilise the evapotranspiration value as the primary indicator of plant water needs, while others define a soil moisture threshold value that determines when to initiate or stop irrigation based on the measured soil moisture in the field. Recent solutions have incorporated AI techniques. However, there is a lack of comprehensive global monitoring of environmental changes affecting plant growth.

Many proposed solutions used the evapotranspiration [5], [6], [7] value as information that describes the plant water needs. Evapotranspiration (ET) is water lost to the atmosphere by two processes-evaporation and transpiration. Evaporation is the loss from open bodies of water, such as lakes and reservoirs, wetlands, bare soil, and snow cover; transpiration is the loss from living plant surfaces[8]. The ET-based solutions use different empirical models to obtain the ET value and various parameters such as rainfall, wind speed, solar radiation, and other weather and air parameters

E-mail address: author 1, author 2, author 3

[9]. The main drawback of these models is that they need exact weather data to give an accurate ET value, which can be difficult to obtain [10], [11].

Other researches have focused on defining a threshold that controls the irrigation [12], [13], [14], which is usually the soil moisture value, where if the current soil moisture value reaches the predefined threshold level, the water pump starts to irrigate the plants. Otherwise, the water pump is off. The choice of soil moisture thresholds can be a challenging point because if the threshold is set too high, the result will be over-irrigation, while setting it too low, will result in under-irrigation and thus production losses [15].

In recent works, researchers have focused on using artificial intelligence (AI) and taking advantage of its ability to solve complex problems in order to predict the exact water amount needed [16], [17], [18], [19].

Although these works may be promising solutions for use as smart irrigation solutions. However, predicting the water amount using the main parameters that affect the plants' irrigation process will be a more meaningful, practical, and forward solution instead of using soil moisture or ET values.

This work aims to propose different ML/DL models that can predict the exact daily water amount the plant needs. The contributions of this paper are presented in the following points:

- Predicting the daily irrigation water amount directly (rather than other relevant values) using various plant environmental conditions during plant evolution (varied from air parameters to soil parameters).
- Studying the impact of data optimization on ML/DL models performance.
- Sensitivity analysis of the proposed models to study their efficiency and robustness.

2. Materials and Methods

The main goal of this work is to predict the daily irrigation water amount that can be affected by plant environmental factors using a given model *M*.

Fig 1 represents the global architecture of the proposed system, which composed mainly of two stages:

- **Training stage:** This stage is applied offline, where we train our model after applying the required data preprocessing and data preparation processes. The data used in this stage are historical data representing a successful experiment of growing a plant with a good irrigation amount decision.
- **Prediction stage:** Once the prediction model is ready, we can use it in the real-time situation from the daily data values captured at the current time to apply the right decision. After a definite time, the current data will be added to the historical data to retrain the model to improve its performance by augmenting the data.

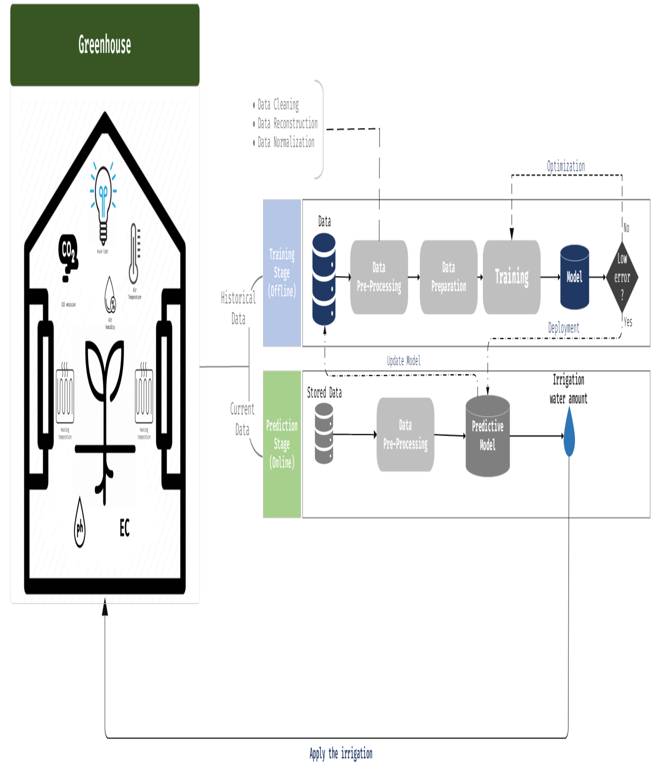


Figure 1. Proposed system process

.

A. Data Description

To predict the amount of irrigation water, it is essential to collect information about the plant environment because it is considered the main factor affecting the irrigation process. For this purpose, we used the Autonomous Greenhouse Competition (AGC) datasets –The first edition– [20]. The AGC is a competition set up in the Netherlands between six teams (5 teams and the experts team) for managing cucumber plant evolution inside the greenhouse (GH). Each team has its own GH with various sensors. Their job is to design and develop machine learning (ML)/deep learning (DL) models that make accurate decisions regarding the plants ' growing conditions (Irrigation/ ventilation/ heating/ CO_2 dosage ... etc.).

The AGC has five dataset collections (teams). Each collection (Team) contains six datasets describing the team's plant evolution environment inside the greenhouse:

- **Crop Management**
- **Greenhouse Climate**

- **VIP**
- **Irrigation**
- **Production**
- **Resources Calculation**

Because the irrigation process can be affected by different plant air environmental changes and soil data [21], we selected the GH Climate and Irrigation datasets of the winning team as the necessary data needed to predict the irrigation water amount.

Table I describes the used data parameters. The GH Climate dataset contains 33133 rows distributed throughout 115 days of growth, where each row corresponds to data recorded for 5 minutes. The irrigation dataset has 115 rows, each row i corresponds to the irrigation information recorded on the day i. The "drain" parameter was used in the AGC competition to determine the net water usage, which is not the case in this situation. For this, we chose to disregard it.

B. Data preprocessing

1) Data cleaning

Data cleaning is a technique used to improve the data quality by detecting and removing errors, and inconsistent and false data [22].

- **Missing values**: Table II shows the number of missing values of the "GH Climate" and "Irrigation" datasets parameters respectively. For the "GH Climate" dataset, we found out that there are 142 missing values for each column –parameter– of a total of 331334 values, which is less than 01% of the total data. For the "Irrigation" dataset, only the water parameter that contains a single value was missed on day 64, which represents 0.87% of the total values (115 rows).
  Because both datasets contain less than 01% missing values of the total data, we ignored these values when we reconstructed the data.

- **Handle outliers**: An outlier is a data point significantly different from the remaining data [23], it's existence in the data may affect the model's performance. For that detecting and handling the outliers is a mandatory step to do. To detect the outliers, we have used the z-score technique.
  Table III shows the number of rows containing outliers accompanying their percentage of the total data. Because the detected outliers carry real data (Not missing values or out of range), the techniques of modifying the outliers may affect the results of our models. So, we decided to study the impact of deleting the outliers on the model's performance by analysing the results with the original data and without the existence of outliers. Thus, we have chosen the outliers detected when *Threshold* = 3 because it contains only 14% of total data (Too small 03% when threshold =4 / Too many when threshold=2).

2) Data reconstruction

Because of the difference in the time intervals between the parameters of the two used datasets, we have reconstructed the used datasets by combining them into a new dataset that contains a unified time interval (Daily interval) by:

- Calculating the daily working time (in minutes) of the "*AssimLight*" parameter.
- Calculating the daily average of the remaining parameters.

The reason for calculating the daily working time of the *"AssimLight"* is because this parameter is categorical in the original dataset (GH Climate dataset), which means that the value of this parameter has only two values: 0%(OFF) or 100%(ON).

Algorithm 1 describe the process of data reconstruction that we have used, The new dataset *Ndata* is the form of a matrix of n columns and m lines, where n corresponds to the number of the parameters used, and m corresponds to the number of samples used which are the growing days. Also, when we make the sum of values, we ignore the missing values that occurred in the original data.

N.B: the value "**288**" mentioned in the algorithm represents the number of rows that form a complete day of data.

3) Data Normalization

- Data scaling: Due to the various type of features used with the different ranges (%, °C, [1-10] for PH,...etc.), data scaling is a crucial step before training the models in order to prevent any model's under-perform caused by the difference in the data ranges. Thus, we have used the *MinMax Scaler* with a feature range of [0-1] to scale our data.

$$X_i' = \frac{X_i - min(X_i)}{max(X_i) - min(X_i)} \tag{1}$$

  Where X_i' corresponds to the scaled value, $min(X_i)$ and $max(X_i)$ are the minimum and maximum values of the parameter i.

- Stationary data: Stationarity means that a stochastic feature does not change even if time changes [24]. Many statistical tests can be used to detect the data stationarity status. One of the most used tests is The augmented Dickey-Fuller (ADF) test. The ADF is the extended version of the simple Dickey-Fuller test, which suppose at first that the data is not stationary (null hypothesis), and then the ADF calculates the p-value. If this value is less than the significant level

TABLE I. Selected Datasets Parameters Description

Dataset	Feature	Description	Unit
GH Climate	Tair	GH air temperature	°C
	RHair	GH relative air humidity	%
	AssimLight	Artificial light used inside the GH	%
	CO_2air	CO_2 concentration in the GH air	ppm
	HumDef	Humidity deficit inside the GH	g/m^2
	Ventwind	Ventilation wind	%
	PipeLow	Temperature of the rail pipe heating on the floor	°C
	PipeGrow	Temperature of the pipe heating on the crop height	°C
Irrigation	pH_Drain	Daily average of drain water PH	[-]
	EC_Drain	Daily average of drain water EC	dS/m
	drain	Daily drain water	l/m^2
	water	Daily irrigation water amount	l/m^2

TABLE II. Missing values for each parameter

Dataset	Feature	Nb. Missing Values
GH Climate	Tair	142
	RHair	143
	AssimLight	142
	CO2air	142
	HumDef	142
	Ventwind	142
	AssimLight	142
	PipeLow	142
	PipeGrow	142
Irrigation	water	1
	EC_Drain	0
	pH_Drain	0

TABLE III. Detected Outliers using z-score test

Threshold	NB. rows	Percentage
02	37	33%
03	16	14%
04	03	03%
05	01	0.87%

(usually 0.5), the ADF reject the null hypothesis and accepts the alternative. The alternative supposes that the time series data is stationary [25]. After passing the ADF test to all the parameters, we find out that:

- *"water", "Pipe Grow Temp", "Daily Pipe Low", "Ventilation wind", "GH Temp" and "Assim Light"* are stationary data.
- The remaining parameters are non-stationary data.

The excising of non-stationarity parameters means that these parameters are related to the time factor. This may invoke a problem with the models' performance. To avoid any possible problem, we need to change this data into stationary data and observe its impact on the models' behaviour. **"Differencing data"** is a common method that changes the non-stationary data into stationary using the following equation:

$$X_i' = X_i - X_{i-1} \tag{2}$$

Where: X_i: Non-stationary data at time i, X_{i-1}: Non-stationary data at time $i-1$, and X_i': Stationary data at time i. Differencing the data can help stabilize

Algorithm 1 Data reconstruction

```
Require: GH_Climate
Require: Irrigation
      Remove unwanted parameters
  GH_Climate ← GH_Climate − {"time_index"}
  Irrigation ← Irrigation − {"drain"}
  Ndata ← {∅}
  for k = 1; i ≤ N do
    for i = 0; i ≤ 114 do
      sum ← 0
      for j = 1; j ≤ 288 do
        current ← GH_Climate[K][i * 288 + j]
        if (k is "AssimLight" and current=100) then
          sum ← sum + 5
        end if
        if (k is not "AssimLight") then
          sum ← sum + current
        end if
      end for
      if (k is "AssimLight) then
        Ndata[k][i] ← sum
      else
        avg ← sum/288
        Ndata[k][i] ← avg
      end if
    end for
    Ndata[k][i] ← avg
  end for
      return (Ndata)
```

the mean of a time series by removing changes in the level of a time series. Therefore eliminating (or reducing) trend and seasonality [26].

C. Data preparation

This section describes the different methods used for splitting the data into inputs and outputs.

1) Simple data preparation

Because the used ML models and MLP model cannot deal with the data as a sequence, we dealt with this data as a discrete problem, meaning that to predict the daily irrigation at day i we gave the model the inputs, which are the environmental change of the plant at the day i with the day index as information that describe the current position of the sequence.

Algorithm 2 describes the process of the simple data preparation.

Algorithm 2 Simple Data Preparation

Require: *data*
Require: *day_index*
$Y \leftarrow data[\text{"}water\text{"}]$ ▷ Output
$X \leftarrow data - \{data[\text{"}water\text{"}]\}$ ▷ All columns except water column (output)
return (X, Y)

2) Time series data preparation

Since the used data includes timestamps, and given the utilization of certain time series Deep Learning DL models in this study, we have introduced a second approach for data preparation for the time series models. This method involves employing the window slide concept to segment the data.

Algorithm 3 outlines the data preparation process utilizing the *window size* parameter. In this context, the output (water amount) on the day i is not only dependent on the inputs (plant's environment parameters) of the day (i), but also on the inputs ranging from $i - window_size$ to the day i.

D. Proposed models

In this work, we proposed different models to predict the irrigation water amount that varying from the ML models to the DL models ending with the advanced time series Models:

- Ada Boost regressor (ABR)
- Extra Tree regressor (ETR)
- Gradient Boost regressor (GBR)
- Multi-layer perceptron (MLP)
- Recurrent Neural Network (RNN)
- Long Short Term Memory (LSTM)

Algorithm 3 Window Sliding Data Preparation

Require: *data*
Require: *window_size*
$Y \leftarrow data[\text{"}water\text{"}]$
$Z \leftarrow data - \{data[\text{"}water\text{"}]\}$
$X \leftarrow \{\emptyset\}$
for $i = 1$to $len(data) - window\_size$ **do**
  $T \leftarrow \{\emptyset\}$
  for $j = i;\ j \leq i + window\_size$ **do**
    $T \leftarrow T \bigcup \{Z[i]\}$
  end for
  $X \leftarrow X \bigcup \{T\}$
end for
return (X, Y)

TABLE IV. ML hyper-parameters' configuration ranges

Model	Parameter	Interval
ABR		
GBR	*n_estimator*	[1-100]
ETR		

E. Evaluation metrics

In this work, we used two metrics to evaluate the performance of the proposed models:

$$Mean\,Absolute\,Error\,(MAE) = \frac{\sum_{i=1}^{n} |Y_i - Y_i'|}{n} \tag{3}$$

$$Root\,Mean\,Squared\,Error\,(RMSE) = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y_i')^2}{n}} \tag{4}$$

Where Y_i corresponds to water amount needs for plant at the day i, and Y_i' represents the predicted output (water amount).

F. Hyper-parameter Tuning

In this section, we will try to identify the hyper-parameters sets for each model that gives the best results. We compared these results using MAE (The hyper-parameter that gives the minimum MAE will be selected).

1) Grid-Search CV

Table IV represents the configuration of the hyper-parameter tuning and the range of possible values of the proposed ML models using the Grid-Search CV method.

2) Bayesian Optimization

Because the used DL models contain many hyper-parameters that need to tune to obtain the best results, we chose to use the Bayesian optimization to find the best hyper-parameters configuration that achieves the best results.

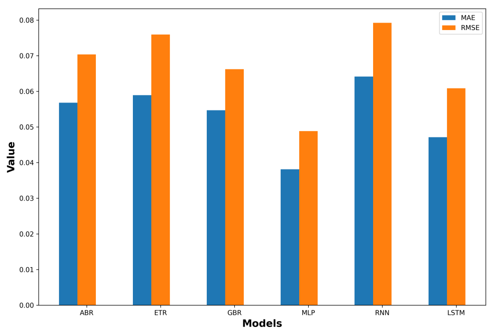


Figure 2. Obtained results without data optimization

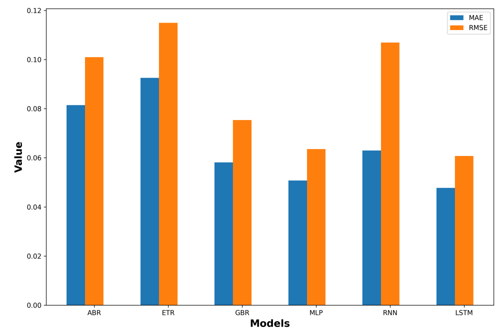


Figure 3. Obtained results after removing outliers.

Table V represents the configuration for the DL models using the Bayesian optimization.

3. Results

A. *Models' performance results*

This section presents the obtained results during different Data optimization phases to evaluate the proposed models' performance. Firstly, we have analyzed the results before applying any data optimization. Then, we have studied the impact of detecting the outliers on the models' behavior and results. Also, because the used data is time-stamped where the irrigation process is applied during all of the growing days and each water quantity could affect , we have studied the impact of stationary/non-stationary data on the proposed models (especially RNN and LSTM). Lastly, both of these optimization techniques have been applied simultaneously to analyze their impact on the models' performance when used together.

1) *Obtained result without any data optimization*

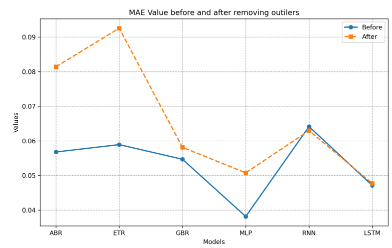
Fig 2 shows the obtained results of MAE and RMSE of the different used models without any data optimization where we can see that the MLP and LSTM models gave the lowest MAE and RMSE values (0.038 and 0.047 respectively for MAE and 0.048 and 0.058 respectively for RMSE).

On the other hand, the RNN model gave the highest MAE and RMSE scores compared to the other models. Also, the $RMSE - MAE$ value (the gap between MAE and RMSE) is the highest, which means that: there is a possible massive error that may occur in this model.

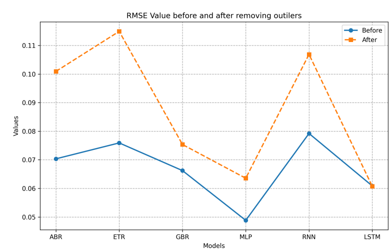
2) *Comparison of results after removing outliers*

Because outliers are one of the main factors that may affect the performance of the models, especially the DL models [27], we decided to study the impact of outliers on the models performance.

Fig 3 shows the obtained results of the ML/DL models after removing the outliers, and Fig 4 shows the difference between these results with the initial results.



(a) Comparison of MAE values



(b) Comparison of RMSE values

Figure 4. Comparison of results before and after removing the outliers

Removing the outliers made the ML models record the worst scores and even worst than the initial results, especially the ABR and ETR models. On the other hand, the DL models gave better results compared to the ML models. Although the LSTM was the best model that gave the lowest MAE and RMSE values among all the models, this did not make any improvement compared to the initial results. For the RNN, we got a decent MAE value. However, the RMSE value was too big (Huge gap between MAE and RMSE), which can be explained by a possible overfitting problem that happened to the RNN model caused by the lack of data by removing the outliers (Data lost).

3) *Comparison of results after differencing data – stationary data–*

Figures 5 and 6 show the results obtained after differencing the data.

TABLE V. DL hyper-parameters' configuration ranges

Layer	Parameter	Range
Fully Connected	*Nb Layers*	[1-10]
	Nb Neurons	[1-500]
	Dropout	[0.1-0.8]
RNN/LSTM	*Nb Layers*	[1-10]
	Nb Neurons	[1-500]
	Dropout	[0.1-0.8]
	Sequence_length/ Window_size	[2-10]
General	*Optimization function*	[*adam, rmsprop, sgd, adadelta, adagard, adamax, nadam*]
	Batch_size	[1-100]
	Epochs	[1-500]
	Layers' Activation function	[*relu, tanh, sigmoid, softmax, softplus, elu, selu*]

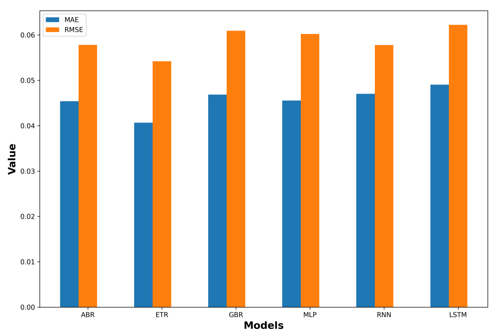
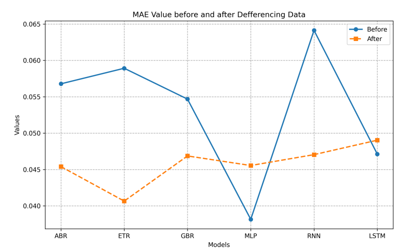
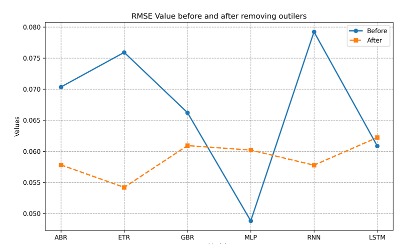


Figure 5. Obtained results after differencing data



(a) Comparison of MAE values



(b) Comparison of RMSE values

Figure 6. Comparison of results before and after differencing the data

These results show a significant improvement concerning all the ML models, especially the ETR method, which gave the lowest results among all the models (0.03 for MAE and 0.03 for RMSE). For the DL model, the RNN model was the only model that took the benefit of differencing the data where it recorded very low MAE and RMSE values compared to the initial results.

4) Comparison of results after applying all Data-based optimization

Figures 7 and 8 show the results after applying all Data-based optimization. The results show that applying both optimization methods did not help at all to improve the model's results. Instead, it gave a much worse score compared to the initial results.

N.B: The Differencing method changes the data values, which could affect the possible outliers. For that, we applied the z-test initially before removing the data.

B. Comparison of models' robustness

Searching for the model that gives the minimum error possible is not enough to deploy any model. Evaluating the model's robustness is a crucial step as well. The models' robustness refers to the ability of a given model to maintain its performance in different conditions [28]. Sensitivity analysis can be used to rank the influence of different hyper-parameters on the model's performance [29], which can help identify the model's robustness by analyzing the distribution of the results during the hyper-parameters tuning. Fig 9 and Table VI represent the result of sensitivity analysis

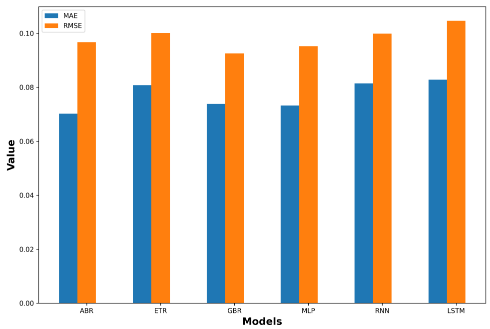
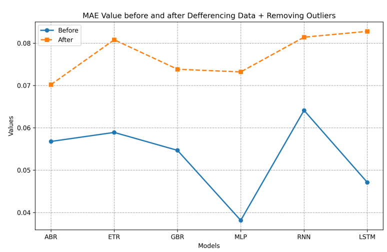
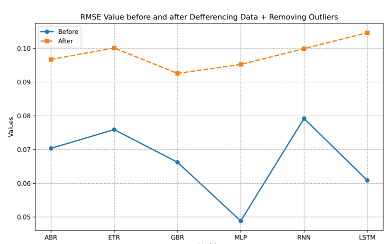


Figure 7. Obtained results after removing outliers + differncing data.



(a) Comparison of MAE values



(b) Comparison of RMSE values

Figure 8. Comparison of results before and after applying all the optimization

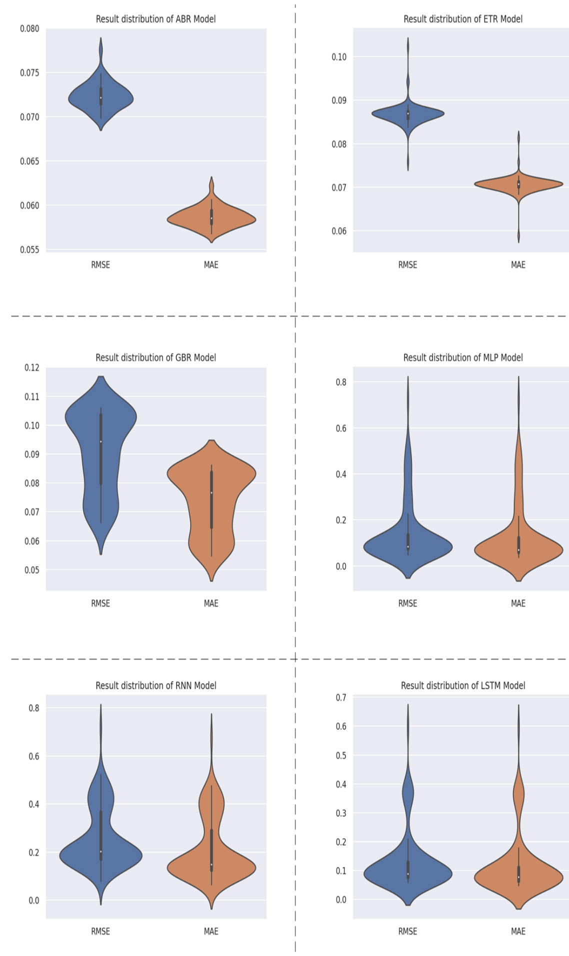


Figure 9. MAE and RMSE results distributions during the hyper-parameters tuning.

of the obtained results of the different hyper-parameters configurations where we have analyzed the distribution of the results to follow every model performance during the different configurations which may help to get an idea about their robustness.

For the ML models, we see that both the ABR and ETR models gave the lowest standard deviation values (Around 0.001), which means that the obtained results of these two models during the different hyperparameters configuration are always close to the mean. This can be explained by the fact that these models have only one parameter that has been tuned, so the deviation of the results will not going to be very wide around the mean. On the other side, the GBR model was the most ML mode sensitive to hyperparameters changes with a notable extensive standard deviation (Around 0.01) compared to ABR and ETR. For the DL models, the interval of the obtained results has notably expanded compared to the ML models where the MLP and RNN gave an MAE and RMSE close to 0.8 in some hyperparameters configurations while the minimum values obtained are too small in some cases where the MLP we got MAE=0.038 which is the best result obtained among all the models, thus has made the standard deviation for these models too huge compared to the ML models, that means the results are skewed to either to the left (Close to the zero) or to the right (Close the max) because the *mean* > *median* in all the DL models that means that these results have a left-skewed where most of the results distributed in the left of the median. This massive difference between the ML and DL models in terms of results distribution can be explained by the fact that we have used a variety of hyperparameters that had changed a lot compared to only one parameter for the ML models, but since almost all results located near to the best results (Close to zero) that may ensure that even these DL models can be robust.

TABLE VI. Statistical analysis of the results of the proposed models

Model	Metric	Min	Max	Mean	Std	Median
ABR	MAE	0.056	0.062	0.058	0.001	0.05
	RMSE	0.069	0.077	0.072	0.001	0.072
ETR	MAE	0.058	0.081	0.07	0.001	0.07
	RMSE	0.075	0.102	0.086	0.002	0.086
GBR	MAE	0.054	0.086	0.073	0.01	0.076
	RMSE	0.066	0.105	0.09	0.013	0.094
MLP	MAE	0.038	0.72	0.127	0.131	0.07
	RMSE	0.048	0.722	0.14	0.128	0.084
RNN	MAE	0.064	0.064	0.21	0.127	0.148
	RMSE	0.079	0.717	0.255	0.123	0.2
LSTM	MAE	0.047	0.593	0.117	0.102	0.076
	RMSE	0.058	0.597	0.13	0.099	0.088

4. Discussion

In general, The MLP was the best model, which gave the best results during all the data optimization phases since the MLP was the model that gave the lowest MAE (0.038 when the no data-based optimization was applied) value with a small error variation magnitude (Small $RMSE-MAE$ value).

The LSTM model also recorded stable results during almost all optimization processes (differencing data) with a low MAE (generally 0.05) and a small *RMSE-MAE* value.

The ML models also gave decent scores with a not too big *RMSE-MAE* value except the GBR model, which was a little bit worse than the other ML models, where this model recorded the highest MAE and RMSE scores.

On the other hand, the RNN model gave the highest MAE and RMSE scores. Although the RNN model recorded the highest error score among the models, its results aren't too poor because even if the RNN gave the worst MAE value (0.08 l/m^3), it still could be considered as a small error margin in the irrigation water amount.

The impact of data-based optimization methods changes from one model to another. Concerning the task of removing the outliers, almost all the models recorded unacceptable reactions to this optimization. This reaction can be attributed to the substantial loss of data, which had previously aided these models in achieving better generalization. Regarding the impact on stationary data, both the ML and RNN models exhibited positive responses to this optimization, yielding noticeably improved results compared to the initial outcomes.

The robustness analysis is a crucial step before deploying any model in real-world situations. The sensitivity analysis of the proposed models shows that the ML models could be more robust than the DL models. However, the possibility of in-time re-training the DL models based on the feedback could be a high advantage that could cover their weakness in terms of robustness.

Acknowledgements

This work is supported by the Algerian Directorate General for Scientific Research and Technological Development (DGRSDT) under the National research projects program in association with the Technical Institute for the Development of Saharan Agronomy (ITDAS) of Biskra.

5. Conclusion

This paper presented various ML/DL models aimed at addressing the issue of water wastage in greenhouse (GH) environments by accurately predicting the daily water amount requirements of plants. The models considered a different set of input parameters, encompassing both air plant environment factors (such as air temperature, humidity, heating temperature) and underground parameters (PH and EC). Results indicate that the Multi-Layer Perceptron (MLP) model outperformed other models, demonstrating superior accuracy with consistent stability throughout all optimization phases.

Concerning robustness, sensitivity analysis revealed the greater robustness of ML models compared to DL models, attributed to the limited number of hyperparameters employed in ML models (typically one parameter). Among DL models, the analysis indicated that the Long-Short Term Memory (LSTM) model exhibited potential robustness, as evidenced by a lower results distribution compared to Recurrent Neural Network (RNN) and MLP models.

However, it is essential to note that losing an extensive amount of data during the data reconstruction phase, particularly from the "GH climate" dataset parameters, may compromise the models' performance. These omitted data could potentially contain crucial information about the dynamic status of the plants throughout the day.

As a future research, a novel technique may be proposed to extract daily fundamental features regarding the plant environment condition, going beyond daily averages. The aim of this approach is to capture essential information directly related to the irrigation process, enhancing the models' overall performance.

References

[1] United Nations Department, "World population prospects 2022: Summary of results," Tech. Rep. UN DESA/POP/2022/TR/NO. 3, 2022.

[2] S. Wanniarachchi and R. Sarukkalige, "A review on evapotranspiration estimation in agricultural water management: Past, present, and future," *Hydrology*, vol. 9, no. 7, 2022. [Online]. Available: https://www.mdpi.com/2306-5338/9/7/123

[3] P. Steduto, J. M. Faurès, J. Hoogeveen, J. Winpenny, and J. Burke, "Coping with water scarcity: an action framework for agriculture and food security," Rome, Tech. Rep., 2012.

[4] K. Chartzoulakis and M. Bertaki, "Sustainable water management in agriculture under climate change," *Agriculture and Agricultural Science Procedia*, vol. 4, pp. 88–98, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210784315000741

[5] M. Dukes, M. Shedd, and S. Davis, "Smart irrigation controllers: Operation of evapotranspiration-based controllers1," *EDIS*, vol. 2009, 04 2009.

[6] S. Davis, M. Dukes, and G. Miller, "Landscape irrigation by evapotranspiration-based irrigation controllers under dry conditions in southwest florida," *Agricultural Water Management*, vol. 96, no. 12, pp. 1828–1836, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377409002340

[7] J. T R, N. S. Reddy, and U. D. Acharya, "Modeling daily reference evapotranspiration from climate variables: Assessment of bagging and boosting regression approaches," *Water Resources Management*, vol. 37, no. 3, pp. 1013–1032, Feb 2023.

[8] R. L. Hanson, "Evapotranspiration and Droughts," https://geochange.er.usgs.gov/sw/changes/natural/et/, 1991, [Accessed 08-Jun-2023].

[9] F. Poyen Phd, d. ghosh, and D. kundu, "Review on different evapotranspiration empirical equations," *International Journal of Advanced Engineering, Management and Science (IJAEMS)*, vol. 2, pp. 17–24, 03 2016.

[10] D. Choudhary, "Methods of evapotranspiration," https://www.researchgate.net/publication/323945435_Methods_of_Evapotranspiration, 03 2018.

[11] A. Mokhtar, N. Al-Ansari, W. El-Ssawy, R. Graf, P. Aghelpour, H. He, S. M. Hafez, and M. Abuarab, "Prediction of irrigation water requirements for green beans-based machine learning algorithm models in arid region," *Water Resources Management*, Mar 2023.

[12] T. Boutraa, A. Akhkha, A. Alshuaibi, and R. Atta, "Evaluation of the effectiveness of an automated irrigation system using wheat crops," *Agriculture and Biology Journal of North America*, vol. 88, pp. 2151–7517, 01 2011.

[13] M. Goodchild, K. Kühn, A. Burek, M. Jenkins, and A. Dutton, "A method for precision closed-loop irrigation using a modified pid control algorithm," *Sensors and Transducers*, vol. 188, pp. 61–68, 05 2015.

[14] J. Gutierrez, J. Villa Medina, A. Nieto-Garibay, and M. Porta-Gandara, "Automated irrigation system using a wireless sensor network and gprs module," *Instrumentation and Measurement, IEEE Transactions on*, vol. 63, pp. 166–176, 01 2014.

[15] T. Kelly, T. Foster, D. M. Schultz, and T. Mieno, "The effect of soil-moisture uncertainty on irrigation water use and farm profits," *Advances in Water Resources*, vol. 154, p. 103982, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0309170821001378

[16] T. Anuşlu, "Smart precision agriculture with autonomous irrigation system using rnn-based techniques," October 2017. [Online]. Available: https://openaccess.mef.edu.tr/handle/20.500.11779/1159

[17] A. Kumar, A. Surendra, H. Mohan, K. Valliappan, and N. Kirthika, "Internet of things based smart irrigation using regression algorithm," *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pp. 1652–1657, 2017.

[18] Y. Huang, H. Jiang, W. feng Wang, W. Wang, and D. Sun, "Soil moisture content prediction model for tea plantations based on SVM optimised by the bald eagle search algorithm," *Cognitive Computation and Systems*, vol. 3, no. 4, pp. 351–360, Sep. 2021.

[19] G. Singh, D. Sharma, A. Goap, S. Sehgal, A. K. Shukla, and S. Kumar, "Machine learning based soil moisture prediction for internet of things based smart irrigation system," *2019 5th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 175–180, 2019.

[20] S. S. Hemming, H. F. de Zwart, A. A. Elings, I. Righini, and A. A. Petropoulou, "Autonomous greenhouse challenge, first edition (2018)," 2019. [Online]. Available: https://data.4tu.nl/articles/dataset/Autonomous_Greenhouse_Challenge_First_Edition_2018_/12717758/1

[21] U. Ahmad, A. Alvino, and S. Marino, "Solar fertigation: A sustainable and smart iot-based irrigation and fertilization system for efficient water and nutrient management," *Agronomy*, vol. 12, no. 5, 2022. [Online]. Available: https://www.mdpi.com/2073-4395/12/5/1012

[22] E. Rahm and H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, pp. 3–13, 01 2000.

[23] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657

[24] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*. Springer New York, 2000.

[25] R. Mushtaq, "Augmented dickey fuller test," *SSRN Electronic Journal*, 2011.

[26] H. Wickham, *ggplot2*. Springer International Publishing, 2016.

[27] A. Khamis, Z. I. ., K. H. ., and A. T. M. ., "The effects of outliers data on neural network performance," *Journal of Applied Sciences*, vol. 5, no. 8, pp. 1394–1398, Jul. 2005.

[28] E. Duesterwald, A. Murthi, G. Venkataraman, M. Sinn, and D. Vijaykeerthy, "Exploring the hyperparameter landscape of adversarial robustness," 2019.

[29] R. Novak, Y. Bahri, D. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: an empirical study," 02 2018.

Hamed LAOUZ is a PhD Student at the university of Biskra (Algeria), specializing in the artificial intelligence field. His research interest focuses on the integration of the AI techniques in the smart-farming sector, especially in the irrigation tasks on the greenhouse climate control.

Soheyb AYAD He is a lecturer in the Computer Science Department at the University of Biskra (Algeria) and a member of the LINFI laboratory. His research interests span various areas, including Networking, Ad hoc Networks, Wireless Sensor Networks, Cloud Computing, Cloud Robotics, Web Services, Semantic Web, Internet of Things, and Predictive Maintenance. Additionally, he has authored several international publications in these fields.

Labib Sadek TERRISSA is Professor in computer science at Biskra University, Algeria. He is also a senior consultant in digitalization project management. After receiving an engineering degree in electronics, he received his Ph.D. in computer engineering in 2006 from Le Havre University, France. His current research interests include Cloud computing, Machine learning, Smart maintenance, and Prognostic and Health Management.

Aicha Nabila BENHARKAT is associate professor at the Computer Science Department (INSA de LYON) since 1992.In this role, she worked on the Integration of heterogeneous databases using Description Logics. Currently, her research interests include the interoperability domain, schema matching techniques in small and large scale, business process as well as the interoperability in service oriented information system and quality and Web services discovery.

Samir MERDACI is an assistant professor in the agronomy faculty at the university of el-oued (Algeria). He is also the head of technical standards department at the technical institute for the development of Saharan agronomy. His research interests primarily revolve around optimizing the utilization of soil and water resources in arid regions.