Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.1, February 2025, pp. 140~145
e-ISSN: 2548-6861

Comparing Machine Learning Algorithms to Enhance Volumetric Water Content Prediction in Low-Cost Soil Moisture Sensor

Iman Setiawan [1*], Mohammad Dahlan Th. Musa [2], Dini Aprilia Afriza [3*], Siti Nur Hafidah [4*]**
* Statistic Study Program, Tadulako University, Palu, Indonesia
** Geophysical Engineering, Tadulako University, Palu, Indonesia
npl.untad@gmail.com [1], ochad1969@gmail.com [2], diniapriliafriza@gmail.com [3], snhafidah07@gmail.com [4]

Article Info

Article history:

Received 2024-11-16
Revised 2024-11-21
Accepted 2024-11-29

Keyword:

Soil Moisture Sensor,
Machine Learning,
Regression,
KNN,
Decision Tree.

ABSTRACT

Measuring soil moisture is possible either with directly using gravimetric test or indirectly using soil moisture sensor. Direct measurements offer accuracy but are not efficient in field measurements. On the other hand, indirect measurement offers remote measurement that will facilitate the user but lacks in accuracy. This research aims to compare and identify the best machine learning model that can improve indirect measurement (soil moisture sensor prediction) using direct measurement (gravimetric test) as a response variable. This research uses linear regression, K-Nearest Neighbours (KNN) and Decision Tree models. The three models were then compared based on Root Mean Square Error (RMSE). The results suggested that KNN (0.02939128) had the smallest RMSE value followed by decision tree (0.05144186) and linear regression model (0.05172371).

I. PENDAHULUAN

Soil volumetric water content (VWC) plays a significant role in the hydrological cycle, particularly in triggering droughts and floods. As a result, proper VWC prediction is essential for effective water resources management as it helps in irrigation planning, increasing crop yields and conserving water. However, traditional methods of measuring VWC, such as taking soil samples or using sensors in the field, can be difficult and expensive, especially for large areas. In addition, complexities in structural properties and relationship with various meteorological factors lead to difficulties in building mathematical models that able to predict soil VWC accurately.[1]

Machine learning (ML) techniques, which use data to make predictions, are gaining popularity in the sensor calibration process as they can estimate VWC more accurately. Machine learning also offers a new way of overcoming the non-linear pattern between the soil dielectric constant of soil moisture sensors and soil moisture content. Thus, developing a more comprehensive model to describe the complex relationship of actual soil water content and sensor measurements that can be applied to various soil types is essential. [2]

Several recent studies have shown that machine learning can help in estimating VWC. First and foremost, [3] calibration of soil moisture sensors using machine learning. The calibration process is then compared between the laboratory and the field by predicting the volumetric water content, the results show that the calibration of the sensor in the field using a linear regression model is better than other machine learning models. Second of all, [4] saw that the decision tree method has a merit benefit to calculate soil moisture since the decision tree is an algorithm that makes a decision tree from given instances. Additionally, utilizing a decision tree offers a chance to save time. Finally, [5] discusses a system that utilizes the KNN algorithm to provide real-time irrigation data to farmers. By considering environmental elements such as soil moisture, temperature, and precipitation forecasts, the system assists farmers in optimizing crop output while minimizing water wastage.

This study will compare three machine learning models—K-Nearest Neighbours (KNN), Linear Regression, and Decision Trees—in predicting VWC across different types of soils and environments. KNN is simple and effective for capturing local data relationships, Linear Regression works well when there is a direct relationship between the data, and Decision Trees are useful for capturing complex data patterns [6], [7], [8]. Using a dataset that includes gravimetric test as direct measurement and sensor as indirect measurements. This study will see which model is most accurate and efficient since indirect methods are beneficial for continuous monitoring in the field, whereas direct gravimetric methods are often reserved for accurate point-in-time measurements and calibration. These findings will help improve water management in agriculture, make irrigation planning more efficient, and promote sustainable agricultural practices.

II. Materials And Methods

The data utilized in this research are primary data from the results of Gravimetric Tests experiments in the field in 2023 using low-cost sensor Capacitive V0.2. The data was then analysed using the help of R software [9]. The explanation of the type and sampling process is as follows:

TABLE 1
CRITERIA FOR DATA COLLECTION

Criteria	Value
Soil ID	Tanah Sekam
Soil Volume (cm^3)	1000
Container Weight (g)	20,0
Dry Soil Weight (g)	600,4
Bulk Density (g cm^{-3})	0,60
Sensor Model	Capacitive V0.2

A. Gravimetric Test

Gravimetric test is a form of direct measurement of soil moisture. The gravimetric method calculates soil moisture content by measuring the weight of a soil sample both before and after drying it in an oven at 105°C for 24 hours. The difference in weight represents the water lost during drying, which allows for calculating the water content. The formula for gravimetric water content (θ_g) is:

$$\theta_g = \frac{W_{wet} - W_{dry}}{W_{dry}}$$

where:
W_{wet} = weight of the wet soil sample
W_{dry} = weight of the soil sample after drying

Although accurate, this method requires a significant amount of time and requires physical sampling, rendering it unsuitable for large-scale or real-time applications [10]

B. Linear Regression

Linear Regression represents the relationship between a dependent variable and one or more independent variables by fitting a straight-line equation to the observed data. In simple Linear Regression, The relationship between the independent variable X and the dependent variable Y is expressed through.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:
β_0 = intercept
β_1 = slope of the line (coefficient of X)
ε = error term

Linear Regression is computationally efficient and interpretable, making it a commonly used baseline for predictive modelling in soil moisture tasks when the relationships are approximately linear [11]

C. K Nearest Neighbour (KNN)

KNN is a non-parametric, instance-based learning method commonly applied to regression tasks. In KNN regression, the prediction for a new data point is calculated by averaging the target values of its k nearest neighbours, which are identified based on Euclidean distance metric. The steps for KNN regression are:

- Step 1: Choose the number of neighbours, k
- Step 2: Determine the distance between the new data point and every point in the training dataset.
- Step 3: Identify the K closest neighbours using the chosen distance metric.
- Step 4: Predict the target value as the average of the target values of the k neighbours

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Where y_i represents the target values of the k closest neighbours.

KNN can capture nonlinear relationships in the data, which makes it useful for soil moisture prediction when local data patterns exist. [12]

D. Decision Tree

A Decision Tree is a supervised learning algorithm effective for handling both classification and regression tasks. In Decision Tree regression, the model splits the data according to feature values, generating "branches" that culminate in "leaf nodes" holding the predicted outcomes. The algorithm recursively splits data at nodes based on criteria (such as minimizing mean squared error), choosing splits that provide the best prediction accuracy. The steps for Decision Tree regression are:

- Step 1: Calculate the variance reduction for each potential split based on the target variable.
- Step 2: Select the split that maximizes variance reduction or minimizes error.
- Step 3: Repeat the splitting process until a stopping criterion is satisfied (e.g., reaching the maximum depth or the minimum number of samples per leaf).
- Step 4: Assign the mean value of target values in each leaf as the predicted value for new data falling into that leaf.

Decision Trees are flexible and handle complex data relationships well but can be prone to overfitting if not properly constrained (e.g., by setting maximum tree depth) [13].

E. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is a widely used metric for assessing the performance of regression models. It is computed as the square root of the average of the squared differences between the predicted values ($\hat{y}_i$) and the actual values (y_i). The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where:
n = number of observations
y_i = measured value
$\hat{y}_i$ = estimated value

RMSE penalizes larger errors more heavily, which is particularly useful for assessing models where larger deviations are critical, such as in soil moisture prediction Lower RMSE values indicate better model performance.[14].

III. RESULT AND DISCUSSION

A. Gravimetric Test

In this study, gravimetric tests were conducted to understand the physical properties of the Chaff Soil, especially in relation to moisture. Some of the important parameters measured include soil volume, container weight, dry soil mass, Bulk Density, wet soil mass, and unprocessed sensor data. Based on the preliminary data, the soil volume used was 1000 cm³, with a container weight of 20.0 grams and a dry soil mass recorded at 995.7 grams. The Bulk Density or dry density of the soil was calculated as 0.06 g/cm³.

This Bulk Density was calculated once only and used as a fixed value at every rise in water volume added to the soil, thus facilitating the calculation of other parameters that depend on soil mass, such as volumetric moisture. The use of this fixed Bulk Density makes the test process more efficient and reduces the repetition of calculations at each stage of adding water. The gravimetric test yielded the following results:

TABLE 2
GRAVIMETRIC RESULTS

Weights of soil and container		m_{wet}	m_{wet}- m_{dry}	θ_g	θ_v	Raw Sensor
Air dry	711	691	90,6	0,15	0,09	418
20	725	705	105	0,17	0,1	395
40	740	720	120	0,2	0,12	392
60	756	736	136	0,23	0,14	354
80	771	751	151	0,25	0,15	357
100	783	763	100	0,15	0,1	318
120	799	779	116	0,18	0,12	368
140	812	792	129	0,19	0,13	365
160	828	808	145	0,22	0,15	352
180	843	823	160	0,24	0,16	377
200	860	840	110	0,15	0,11	362
220	873	853	123	0,17	0,12	340
240	886	866	136	0,19	0,14	310
260	901	881	151	0,21	0,15	275
280	911	891	161	0,22	0,16	274
300	926	906	119	0,15	0,12	228
320	952	932	145	0,18	0,14	216
340	966	946	159	0,2	0,16	202
360	972	952	165	0,21	0,16	201
380	989	969	182	0,23	0,18	194
400	1006	986	129	0,15	0,13	186
420	1026	1006	149	0,17	0,15	182
440	1041	1021	164	0,19	0,16	180
460	1054	1034	177	0,21	0,18	176
480	1070	1050	193	0,23	0,19	174
500	1087	1067	140	0,15	0,14	172
520	1103	1083	156	0,17	0,16	170
540	1123	1103	176	0,19	0,18	169
560	1135	1115	188	0,2	0,19	167
580	1151	1131	204	0,22	0,2	166
600	1166	1146	150	0,15	0,15	165

Table 2 shows the measurement data of soil moisture in the container at different levels of water addition, ranging from dry to wet conditions. The first column, "Soil + Container Weights," records the combined weight of the soil and the container in grams after gradual water addition. For example, under air-dry conditions, the weight of the soil and container is noted as 711 grams and increases to 1087 grams when the amount of water reaches the maximum level in the last row.

The "m_{wet}" column indicates the weight of the soil under wet conditions. For example, the weight of wet soil in air-dry conditions is 691 grams, which increases to 1067 grams at the maximum water level. The "m_{wet} - m_{dry}" column represents the difference between the wet soil weight and the dry soil weight, reflecting the amount of water in the soil at each moisture level. In air-dry conditions, this value is 90.62 grams, which increases to 193.31 grams in rows with a total weight of 1087 grams.

The column "θg" (gravimetric moisture) shows the ratio of water mass to dry soil mass in grams per gram (g g^{-1}). Under air-dry conditions, the gravimetric moisture is 0.1509 g g^{-1} and reaches 0.2257 g g^{-1} at the maximum water point. The column "θv" (volumetric moisture) describes the amount of water per unit volume of soil in cubic meters of water per cubic meter of soil (m^3 m^{-3}), which is important for measuring water availability for plants as well as irrigation needs. Under air-dry conditions, the volumetric moisture value was recorded as 0.0906 m^3 m^{-3} and increased to 0.2039 m^3 m^{-3} at the highest moisture content.

The last column is "Sensor Measurements (RAW)," which records the raw data from the soil moisture sensor at various moisture levels. The sensor reads a value of 418 in dry air conditions and decreases to 172 at the highest moisture content, indicating changes in soil moisture levels that can be monitored to maintain optimal conditions for plant growth. Overall, this table is useful for analysing crop water requirements, determining field capacity, and evaluating optimal soil irrigation conditions.

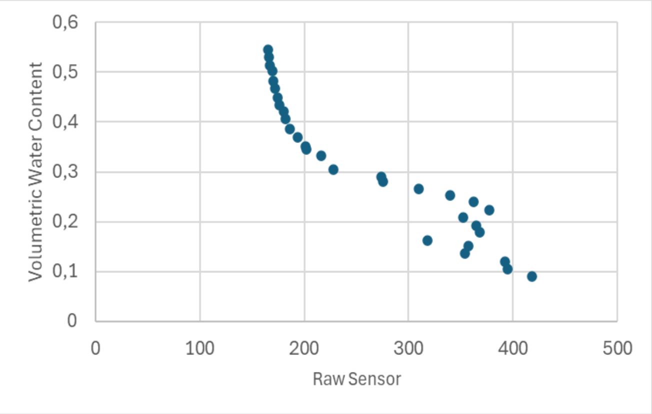

Figure 1 Scatter Plot of Raw Sensor Data and VWC

Figure 1 depicts the relationship between sensor raw data (X) and volumetric water content (Y). It shows a clear trend where the water content decreases as the sensor raw data increases, but the relationship might not be strictly linear. The sensor readings range from 165 to 418 and water content ranges from 0.0906 to 0.5456. The dataset size is relatively small, with 31 data points. The results of the volumetric water content measurement and sensor raw data model are as follows.

B. Linear Regression

In linear regression analysis, two important metrics often used to evaluate model performance are R-squared (R^2) and Root Mean Square Error (RMSE).

TABLE 3
LINEAR REGRESSION MODEL EVALUATION

R-squared	RMSE	
	Training	Testing
0,8845452	0.04353125	0,05172371

Table 3 shows that R-squared, which in this case is 0.8845452, Is a metric that reflects how effectively the regression model explains the variability in the data. R^2 values range from 0 to 1, with higher values closer to 1 suggesting that the model accounts for most of the variation in the data. In this case, the R^2 of 0.8845452 shows that approximately 88.45% of the variation in the dependent variable is accounted for by the constructed linear regression model. This indicates the model has a good fit with the data.

On the other hand, the RMSE of 0.05172371 measures The mean difference between the model's predicted value and the actual value using data testing. A lower RMSE indicates that the model is more accurate in predicting the true values. In this context, the RMSE of 0.05172371 indicates that the average prediction error of the model is about 0.05, which is considered very small, indicating that the linear regression model provides very accurate results.

C. K Nearest Neighbour

In K-Nearest Neighbours (KNN) model analysis, it is important to understand how the number of neighbours (K) parameter affects model performance. One way to evaluate model performance is by using the RMSE, which measures the average error between the predicted value and the true value. The graph presented shows the relationship between the K value and the resulting RMSE, providing insight into how the choice of K can impact the accuracy of the model.

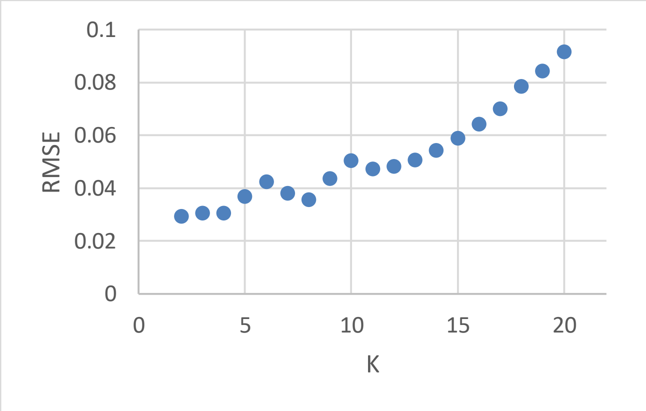

Figure 2 Number of Neighbours(K) Simulation

Figure 2 shows that the number of neighbours (referred to as K) in the KNN model ranges from 0 to 20, while the RMSE is plotted on the Y-axis. The dots on the graph represent the RMSE for a given K value, starting with a lower RMSE for small K values and generally increasing as K increases. This trend shows that as K increases, the RMSE tends to increase, indicating that a higher number of neighbours may cause the model to be too smooth, resulting in a poorer fit to the data. Thus, the optimal K value seems to be at K=2, where the RMSE reaches a minimum. Beyond this value, increasing K seems to lead to higher error rates, indicating that the predictive performance of the model deteriorates.

TABLE 4
KNN MODEL EVALUATION

	RMSE
Training	0.01974099
Testing	0.02939128

Table 4 shows the gap from predicted value and the actual data using data training and testing. The RMSE at 0.01974099 indicates how good the KNN model is in predicting data training. Furthermore, The RMSE at 0.02939128 also shows how effective the KNN model is in predicting the testing dataset.

D. Decision Tree

In decision analysis using decision trees, it is important to understand how splitting criteria are used to predict outcomes based on variable values.

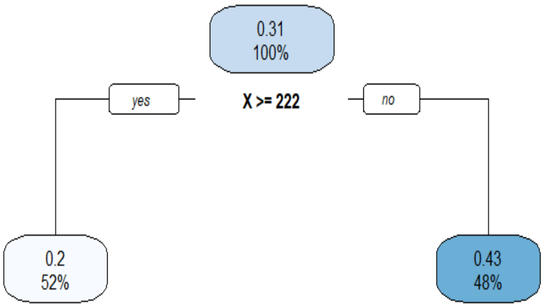

Figure 3 Decision Tree Diagram

Figure 2 shows a decision tree structure that illustrates the classification process based on the variable X, with a threshold value of 222. At the top of the tree, it can be seen that if the value of X is greater than or equal to 222, then the classification leads to the branch "yes," which results in a value of 0.31 with 100% of the data meeting that condition. On the other hand, if the X value is less than 222, the classification leads to the "no" branch. From the "no," branch, the RMSE value obtained is 0.43, with a proportion of 48% of the total data.

Furthermore, in the "yes" branch, there was an RMSE value of 0.2, with a proportion of 52% of the data indicating that the model gave better predictions in this group compared to the "no" group. This shows that the split based on the threshold value of X = 222 is effective in distinguishing the two different outcome groups, with the "yes" group performing better with a lower RMSE. This tree yielded an RMSE of 0.05144186.

TABLE 5
DECISION TREE EVALUATION

	RMSE
Training	0.06941745
Testing	0.05144186

Table 5 illustrates the discrepancy between the predicted values and the actual data for both the training and testing datasets. The RMSE value of 0.06941745 demonstrates the effectiveness of the decision tree model in predicting the training data. Additionally, the RMSE value of 0.05144186 highlights the model's capability in accurately predicting the testing dataset

E. Low-Cost Sensor Calibration

The sensor calibration process is intended to configure the sensor output numbers so that they can be easily understood and easy to program. In addition, it is expected that the modelling machine can accurately predict the pattern of direct measurement data (gravimetric test) using the best machine learning model. Therefore, it is necessary to evaluate the performance of the three models using training and testing datasets. It is necessary to study the best model that can capture the direct measurement data pattern in the training data with the smallest RMSE value in the testing data.

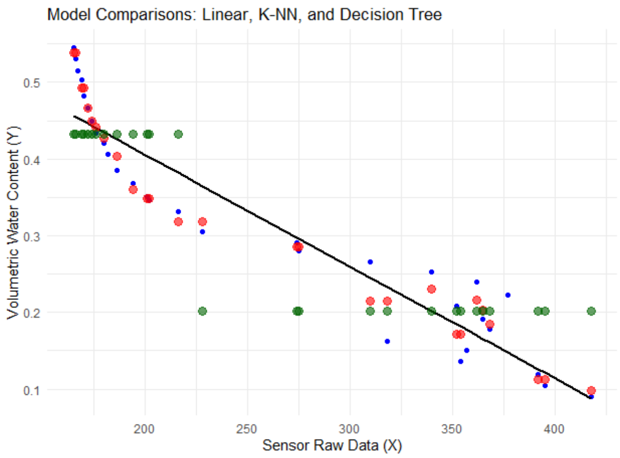

Figure 4 Compares three model prediction in R

Figure 4 shows a chart that compares the performance of three regression models Linear Regression, K-NN and Decision Tree based on their ability to fit training data. The Decision Tree, shown with green points, does not effectively capture the non-linear trends in the data, with the lowest RMSE value at 0.06941745. Linear Regression, represented by the red line, provides a simple linear fit. However, it struggles to perform well due to the non-linear relationship between the variables, resulting in higher error rates. It achieves a lower RMSE compared to K-NN.

K-NN, represented by orange points, adapts to the local structure of the data and excels in capturing complex patterns. It achieves the lowest RMSE among the three models, demonstrating its strength in handling small datasets with non-linear trends. Overall, K-NN outperforms the other models as it does not assume a global relationship between variables and instead relies on local patterns, making it particularly effective for datasets with non-linear relationships and limited size. Furthermore, model evaluation on testing dataset is as follows:

TABLE 6
BEST MODEL EVALUATION ON TESTING DATASET

Model	**RMSE**
Linear Regression	0,05172371
K-NN	0,02939128
Decision Tree	0,05144186

Table 6 shows that of the three models tested using the testing dataset, K-NN provides the most accurate prediction results, with an RMSE value of 0.02939128. A smaller RMSE value indicates that the model has a lower prediction error, making it more effective in predicting data compared to other models.

Although the Decision Tree and Linear Regression models have similar RMSE values, they still have higher errors than the K-NN. These results show that K-NN is the best model for this data, while Linear Regression and Decision Tree perform less optimally in terms of prediction accuracy.

Based on that result, the calibration process of the low-cost sensor begins with programming the KNN model of the Arduino. The code consists of a couple of steps like reading the sensor data, calculating distances and finally making a prediction with an average value of the three nearest neighbours. The code for calculating distances and make prediction is given as follow:

```
float calculateDistance(float x1, float x2) {
  return abs(x1 - x2); }
float predictKNN(float x) {
  float distances[dataPoints];
  int nearestK[2] = {0, 0};  // Initialize
for (int i = 0; i < dataPoints; i++) {
    distances[i] = calculateDistance(x,
X_train[i]);
}
for (int i = 0; i < dataPoints; i++) {
  for (int j = 0; j < 2; j++) {
    if (distances[i] < distances[nearestK [j]]) {
        nearestK[j] = i;
        break;}
   }
 }
float y_pred = (Y_train[nearestK [0]] +
Y_train[nearestK[1]]) / 2;
  return y_pred;}
```

F. Discussion

The use of machine learning in research is to predict volumetric water content (VWC) while transforming data and ease of integration of the coding process in Arduino if using many sensors simultaneously locally on the sensor device.

Initially, VWC measurements using sensor measurements are in the range of 165 - 418 with the smaller the sensor data, the greater the VWC which is difficult to read and results in misunderstanding of the amount of water in the soil for users. For example, when the sensor output at time t is 170 and then after watering at time $t+1$ becomes 260, it cannot be directly interpreted that the increase in water volume is 90. On the contrary, after utilizing machine learning such as KNN, the measurement results in the range of 0.09 - 0.5 are obtained which we can directly interpret the increase.

Many researchers have tried to deal with this by converting it into a proportion. However, the output of low-cost sensors such as capacitive has a different range, so the process of forming a percentage range needs to be done for each sensor. In the application of watering systems using multiple sensors, this becomes complex and requires careful integration into the Arduino system. However, after using KNN, even though using several sensors we can simplify the coding process in Arduino to get the same range of values.

Therefore, the use of machine learning can be used as an alternative to smoothing the volumetric water content numbers on cheap sensors other than using percentages. In addition, the use of machine learning offers ease of coding in watering installations using more than 1 sensor.

V. CONCLUSION

KNN saw the most accurate prediction and lower prediction error. It has 0.02939128 RMSE value, confirming this model as a more effective way to predict soil moisture compared to the other models.

ACKNOWLEDGEMENT

We affirm and deliver our gratitude toward both LPPM and FMIPA UNTAD for funding this research and suggesting pivotal suggestions and expertise that have a huge impact on this publication.

REFERENCES

[1] K. Nath, P. C. Nayak, and K. S. Kasiviswanathan, “Soil volumetric water content prediction using unique hybrid deep learning algorithm,” Neural Comput Appl, vol. 36, no. 26, pp. 16503–16525, Sep. 2024, doi: 10.1007/S00521-024-09991-6.

[2] H. Wan, H. Qi, and S. Shang, “Estimating soil water and salt contents from field measurements with time domain reflectometry using machine learning algorithms,” Agric Water Manag, vol. 285, Jul. 2023, doi: 10.1016/j.agwat.2023.108364.

[3] S. Adla, F. Bruckmaier, L. F. Arias-Rodriguez, S. Tripathi, S. Pande, and M. Disse, “Impact of calibrating a low-cost capacitance-based soil moisture sensor on AquaCrop model performance,” J Environ Manage, vol. 353, Feb. 2024, doi: 10.1016/j.jenvman.2024.120248.

[4] E. Pekel, “Estimation of soil moisture using decision tree regression,” Theor Appl Climatol, vol. 139, no. 3–4, pp. 1111–1119, Feb. 2020, doi: 10.1007/S00704-019-03048-8/METRICS.

[5] M. Thenmozhi, P. Subbiah, J. Justina Michael, and R. Theobard, “Optimizing Agricultural Irrigation through Real-Time Data Prediction,” Proceedings of the 2nd IEEE International Conference on Networking and Communications 2024, ICNWC 2024, 2024, doi: 10.1109/ICNWC60771.2024.10537511.

[6] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications,” Journal of Big Data 2024 11:1, vol. 11, no. 1, pp. 1–55, Aug. 2024, doi: 10.1186/S40537-024-00973-Y.

[7] P. K. Syriopoulos, N. G. Kalampalikis, S. B. Kotsiantis, and M. N. Vrahatis, “kNN Classification: a review,” Ann Math Artif Intell, pp. 1–33, Sep. 2023, doi: 10.1007/S10472-023-09882-X/METRICS.

[8] R. G. McClarren, “Decision Trees and Random Forests for Regression and Classification,” Machine Learning for Engineers, pp. 55–82, 2021, doi: 10.1007/978-3-030-70388-2_3.

[9] R Core Team, “R: A Language and Environment for Statistical Computing,” 2023, Vienna, Austria. [Online]. Available: https://www.r-project.org/

[10] A. Trivedi, M. Gandhi, N. Nandeha, Y. Rajwade, and V. R. Rao, “Sensors of Soil Moisture Measurement,” 2023. [Online]. Available: https://www.researchgate.net/publication/375285282

[11] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, and P. Finke, “Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran,” Geoderma, vol. 338, pp. 445–452, Mar. 2019, doi: 10.1016/J.GEODERMA.2018.09.006.

[12] K.M.Annammal, S.Porkodi, V.Jerald Abishek, C.Thomas Abraham, and J.Dhana Babu, “Automated Irrigation for Smart Gardening based on IOT using sensors,” international journal of engineering technology and management sciences, pp. 117–122, Sep. 2022, doi: 10.46647/IJETMS.2022.V06I05.016.

[13] B. Babayigit and B. Buyukpatpat, “Comparison of Machine Learning Regression Models for the Prediction of Soil Moisture with the use of Internet of Things Irrigation System Data,” Journal of Institute of Science and Technology, vol. 37, no. 3, pp. 479–487, Dec. 2021.

[14] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” Geosci Model Dev, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/GMD-15-5481-2022.