Abhishek Sharma
2Arvind Kumar Shukla
3Kolli Himantha Rao
4Manish Singh
5Elangovan Muniyandy
6Sridhar S

Machine learning and Sensor-Cloud Based Precision Agriculture for Intelligent Water Management for Enhanced Crop Productivity

Abstract: The combination of Machine Learning algorithms with Internet of Things devices is emerging as an effective solution to redefining precision agriculture for better water management and crop cultivation. The purpose of this study is to use different learning models, such as Artificial Neural Networks, Support Vector Machines, Decision Trees, and Random Forest, to predict the irrigation need based on real-world sensor data. To retrieve the output variable, which is the irrigation requirement, data from temperature, soil moisture, water level, and humidity sensors that are available in a tomato cultivation facility are used. The dataset consists of 3422 readings, which are split into training and testing sets. A designated percentage of 70% is used for training the models, while the remaining 30% are used to test the outcomes. As the results of the study show, the ANN model is the most accurate predictor of irrigation need with the classification rate of 97.6%, followed by SVM, DT, and RF with 95.4%, 91.3%, and 88.9% correspondingly. The differences in the outcomes are demonstrated in confusion matrices, which identify the classification of the cases and indicate the percentage of correct predictions. Evidence of the predictive power of ML models implies that farmers can independently determine when to activate the pump when real-world data serve as the input. Additionally, the ability to collect real-world data using IoT sensors is beneficial, as it empowers farmers with up-to-date information to make a timely decision about pump activation. The limitations are associated with the type of crop and agricultural facility and, for this reason, future studies may investigate the generality of the conclusion with regard to other crop types and facilities.

Keywords: precision agriculture, machine learning, IoT technology, water management, crop productivity

I. INTRODUCTION

Precision agriculture is a modern paradigm of incorporating advanced technologies, such as Machine Learning and Internet of Things, and enhancing farming operations [1]–[3]. The exponential growth of the population combined with resource constraints and environmental issues indicate that innovative solutions are required to manage agricultural processes. Water management is one of the central aspects of precision agriculture: uneven irrigation often results in significant inefficiencies and waste of resources. Although the moisture of the soil and crop needs vary depending on the type of plant and external conditions, traditional form of irrigation do not allow due precision of water allocation [4]–[6]. Recognizing the potential waste of valuable resources, timely switching on of water supply, or, in contrast, late irrigation, it becomes apparent that intelligent water management is required. IoT technology enables farmers to collect real-time data using sensors embedded in the soil. Farmers can observe the level of moisture, temperature fluctuations, and the health of the crop. ML algorithms, in turn, analyze...
patterns observed in the previously collected data and make some predictions, automate decision-making, and optimize resource allocation [7], [8]. The aim of this project is to examine the potential of IoT technology, ML algorithms, and water management as the three areas in which precision agriculture is frequently utilized. The study will summarize their benefits and challenges and arrive at suggestions which will be useful for practitioners in the field and support informed decision-making.

II. LITERATURE REVIEW

Precision agriculture, a diverse data-driven farming approach, has experienced significant growth over the past years. The major contribution to this development was done by machine learning tools. In a literature review, the author discusses how the logistic regression is applied in precision agriculture and the benefits and drawbacks of such an application. The purpose of the review is to investigate the application of the logistic regression in different areas, such as crop classification, disease detection, and yield prediction as well as usage for a decision support system [9], [10].

The main application of the logistic regression is crop classification. The authors of the study further explore the classification of crops based on the spectral data, received from remote sensors and collected by hyperspectral imaging. The algorithm was trained on the data containing the unique spectral particularities of each type of fields by the researchers to achieve the high percentage of correct classification of the fruit. This, therefore, allows farmers to track the state of their fields remotely and to apply the most suitable precision disease control. The researchers also assume that “sensor data collected on environmental conditions, crop type, and such plant parameters as value of the leaf area index variety could be used in predicting disease activity and post-inoculation disease”. Hence, the prediction system could be implemented with the logistic regression, so the farmer would receive a timely warning about the possibility of the disease to spread. The same data could be used for the pest population modeling, and if the logistic regression was applied to them, farmers would always know if they need to apply a pesticide this year or not [11]–[13].

Thus, many applications for logistic regression in precision agriculture exist. They include crop classification and disease spread prediction, and the promising applications involve crop yield prediction and decision support system design. The algorithm can be deemed as generally useful for these applications, although one must always remember its high sensitivity to multicollinearity, potential model overfitting, and risk of non-linear relationships. Still, the disadvantage of these qualities can be avoided by appropriate data preparation and understanding. The benefits of using logistic regression are primarily related to the basics described above, as well as the method’s simplicity [14]–[16]. It is vastly easier to employ and understand its results compared to the application of advanced machine learning algorithms, such as neural networks or support vector machines. Logistic regression results are simple because they rely on interpreting the logistic function of the model’s coefficients, thus allowing the farmer or the specialist to understand how predictors influence the probability of the outcome. In other words, the approach is highly beneficial for precision agriculture decision-making as it provides stakeholders with clear information on their impact and encourages simple, actionable message delivery. Logistic regression is also advantageous because it works well with small to medium datasets, an important quality for precision agriculture where aggregating significant data may be costly or challenging in some areas. Thus, to avoid multicollinearity, apply such algorithms and methods that support feature selection or use regularization. As a result, the model will be more robust and generalized. The simplicity and practical significance of this approach are the reasons for its being one of the most common methods in many fields and applications, including precision agriculture. [17]–[19].

Regarding whether logistic regression seems useful in practice for facilitating decision-making in the context of precision agriculture, it seems feasible as the technique offers one of the most interpretable and effective solutions to the classification problems; thus, it might have a certain potential contribution to the agricultural decision-making in the long run. However, more research has to be done in order to extend the overall applicability of logistic regression in more complex fields that are related to agriculture. For example, it is possible to focus on the incorporation of domain-specific knowledge and features into logistic regression to enhance its usability in the agricultural domain. [20]–[22].

It is suggested to focus on the use of two models and identify the approach of their application to the agricultural dataset. This method will help to ensure that the technique is used in the most suitable and context-related manner,
along with identifying the key issues typical for agricultural datasets but not other categories. Another recommendation is to concentrate on the use of ensemble techniques along with logistic regression, as they are expected to offer more advanced tools in use, with their combination depending [23–25]. The logistic regression can be used for decision support in precision agriculture and in combination with other similar algorithms can improve their combined power. So, one can state that knowing its pros and cons, logistic regression can be considered to be one of the most powerful ways to improve the agricultural prospects, diminish environmental risks and make agriculture more sustainable, if used carefully.

III. METHODOLOGY

Various sensors, namely temperature sensors, soil moisture sensors, water level sensors, humidity sensors play a central role in collecting real-time data about plant’s water requirement. These sensors provide critical feedback for understanding the environmental conditions affecting plants growth and facilitate an accurate water management system. The developed IoT device is interfacing these sensed parameters.

The IoT device serves as the interface for the sensors and enables the measuring and transmitting of the feedback to the controller. On analyzing the transferred data in real-time, the information about the water requirements of the plant are constantly being monitored.

![Figure 1. Methodology of the proposed system](image)

The interfaced controller receives the real-time data from the sensors. The controller then performs the important function of analyzing the received data and needs to understand the real-time IoT system’s decision-making capabilities. It has to assess and determine the actual real-time information about the moisture levels of the soil, temperature of the soil, humidity of the soil, water levels for different crops. When the current received sensed data indicate that the soil moisture has gone below the required rate, the pumping system is activated. These decided actions need to happen very fast to control the amount of water delivered to the soil.

One main ICT based advantage of the designed IoT system is that it is able to respond quickly on getting a meaningful shift form the sensor data. If for any reason, the soil’s moisture moves below the required level, then the pumping system is activated. This personifies the automated response system developed for data collected from sensors interfaced for the desired crops only.

In addition to interfacing different sensors, the complete recorded sensory data from the sensors is also transmitted to the cloud on one side. Data stored in the cloud will permit the farmers, agronomist, and researches to monitor the crop real-time data on the field. Meanwhile, the Information will be stored in one data store and most importantly, it enables stored data analysis after data of the complete year is collected.
IV. NEED FOR MACHINE LEARNING MODELS

Since along with the developed IoT system, it is necessary to provide an advanced prediction of the water level requirements and contingency measures for the pump operation in case of the functioning of the developed system failure. To do so, a Machine Learning model would be integrated into the system. The purpose of the Machine Learning model is to predict water requirements well in advance based on historical sensor data and corresponding pump operations. By analysing the patterns and relationships within the data, the purpose of the model is to forecast the future water requirements of the crops highly accurately. This allows for proactive decision-making, permitting irrigation schedules to be adjusted in advance bewowing optimal water utilisation and assistance to the health of the crops.

At the same time, the Machine Learning model serves as contingency measures for the developed system’s failure and can prevent it providing pump operations regardless of the functioning of the IoT system controller. The model is trained on historical sensor data and corresponding pump operations. The patterns within these data are critical as they can signal about the needs for irrigation. Thus, in case if the controller of the developed system does not work, the Machine Learning model would still be able to manage the pump. The model independently turns on/off the pump based on the sensor reading at the moment. Thus, the system includes a reliable internal measure that can manage the pump in any situation independently. Moreover, the more environmental data the model can process and the more operations of the pump it can analyse, the more effective it manages pump operations, minimizing not only water use but the load on the pump.

A. Various ML model used in the research

The current research utilizes the use of various Machine Learning models including Artificial Neural Networks, Random Forest, Decision Trees and Support Vector Machines. The purpose is to enhance the accuracy and efficiency of water management in agriculture. Artificial Neural Networks are computational models that attempt to simulate the biological neural networks in human brain. The model comprised of multiple interconnected nodes which are presented in layers. In every node, a series of simple numerical operation is performed. ANN is best in the field of pattern recognition and nonlinear data modeling to predict the complex relationships present in agricultural data. Through extensive training, the data of historical sensor data and corresponding water requirement is inputted into the ANN which learns to predict irrigation needs in the future. Random Forest is an ensemble learning method that trains a multitude of decision trees and outputs the class mode or the mean prediction. The key advantage of this method is that it is not prone to overfitting and is effective in high-dimensional input problems. In precision agriculture, Random Forest is used to analyze the vast sensor data to determine the significant features and patterns which influence the water requirement in farming. The use of the mean prediction or class mode of the invididual decision trees in Random Forest increases the accuracy of water management prediction.

Decision Trees is a grid-based machine learning model that recursive partition the input data space in the feature sub-space based on the attributes of input features. The partition is associated with a decision node. The final outcome is the leaf nodes of the decision trees. The key features of DTs is its ease of interpretation and provision of deep insight to the decision process. In agriculture, the DT can partition the optimal irrigation data based on the environmental parameter of the soil moisture. Such simple and intuitive nature of the tree structure can aid farmers in making the decision in this regard. Support Vector Machines is a supervised method utilize for classification and regression analysis. This method finds the hyperplane which best separate the multiple classes or maximises the data point margin. SVM is best applied in high-dimensionally input datasets and its ability to generalize the data to the test data. In precision agriculture, it can be used for the classification and irrigation prediction based on the soil moisture and other sensors data.

V. PREPROCESSING AND FEATURE ENGINEERING

The first step of the IoT system operation takes place when the system is turned on, and the sensors are deployed. At this point, various data about the environment is collected, including temperature, soil moisture, water level, and humidity. These data are then transferred to the cloud system where the data is stored and confronted.

The database contains in total 3422 rows of data. These pieces of data offer descriptions of a wide spectrum of environmental conditions and paired descriptions of activation of the pump. It is arranged to best suit the Machine
A substantial number of preprocessing steps and feature extraction techniques have been applied to the dataset to ensure its quality and suitability for the Machine Learning task in precision agriculture. A number of steps of cleaning, standardization, normalization, and feature extraction have been performed in that way. The first step was data cleaning, which included addressing the presence of inconsistencies, outliers, or missing values in the collected dataset. A number of techniques, such as data imputation and exclusion of the entries, which had insufficient or invalid data, has been applied. The action resulted in cleaned data, which value has been improved. The second important step was standardization and normalization, which allows fitting the variables to a comparable magnitude. The large difference in magnitude for different features implies that those features with a larger magnitude will dominate the model training. Through standardization and normalization, all ranges of features were ensured to contribute in equal measure to the model learning. The following important step of preprocessing that has been applied and is widely applied in the assignment in question is the feature extraction, or the extraction of relevant information from the raw sensor data.

The step involved the identification and selection of sensor features that can be most useful to predict the target variable, such as the need of the pump to be activated. The first type of feature extraction that was used in the assignment involved calculating mean, median, standard deviation, and variance of the sensor readings during certain time intervals. Additionally, the second part of the assignment made a significant contribution by considering domain knowledge. Thus, depending on the physiology of different crops and their varying requirements in water, the most important features were selected. For example, soil moisture levels during different growth stages and temperature fluctuations reflecting water stress were used in the assignment.

There are several steps involved in preparing a dataset for fitting machine learning models and predicting pump activation based on sensor readings. For instance, missing values can be imputed and inconsistent data can be preprocessed, including filling missing values, encoding categorical strings, and normalizing data. Since using all known data may be computationally costly and less interpretable, feature selection tools such as PCA or RFE may be employed to use fewer features and to avoid redundancies or noisiness of input data.

Additionally, in addition to the numerical features derived from the sensors, there can be categorical features in a dataset that cannot be directly put into a model. Examples are strings with a geographic area, crop type, or soil type, which can influence the water demands and pump activation. These variables are often transformed into numerical variables with a method such as one-hot encoding or label encoding.

Once the dataset is clean and features are defined, it is split into training and testing sets at e.g. 70/30 ratio. In the cases examined below, the training set is employed for fitting machine learning models using one of the following supervised learning algorithms: ANN, RF, DT or SVS. The models learn patterns to discover the relationships between the input variables – the characteristics of the soil and Africa precipitation and temperature – and the output variable – namely, one means the pump does not need to be activated, and 0 means the pump needs to be activated. Finally the models are evaluated in terms of their prediction accuracy by running them on the test set using metrics such as accuracy, precision, recall, and f1-score to compare their effectiveness.

VI. RESULT AND DISCUSSION

The table 1 presents the sensor readings recorded from morning to afternoon, and the pump operation status. These features, such as temperature, soil moisture, water level, and humidity readings, are provided to the machines as the input for their ML models. While pump operation is the target variable, it is used for training the models to understand the connection between the provided features and irrigation. If the machine is trained on such a dataset, it will be able to predict whether the pump should be running on the previously recorded parameters.
TABLE I. SENSOR READINGS AND ACTUATOR RESPONSE

<table>
<thead>
<tr>
<th>Time</th>
<th>Temperature (°C)</th>
<th>Soil Moisture (%)</th>
<th>Water Level (cm)</th>
<th>Humidity (%)</th>
<th>Pump Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00 AM</td>
<td>22</td>
<td>40</td>
<td>15</td>
<td>60</td>
<td>Off</td>
</tr>
<tr>
<td>7:30 AM</td>
<td>23</td>
<td>42</td>
<td>14</td>
<td>62</td>
<td>Off</td>
</tr>
<tr>
<td>8:00 AM</td>
<td>24</td>
<td>45</td>
<td>13</td>
<td>63</td>
<td>Off</td>
</tr>
<tr>
<td>8:30 AM</td>
<td>25</td>
<td>47</td>
<td>12</td>
<td>65</td>
<td>On</td>
</tr>
<tr>
<td>9:00 AM</td>
<td>26</td>
<td>49</td>
<td>11</td>
<td>67</td>
<td>On</td>
</tr>
<tr>
<td>9:30 AM</td>
<td>27</td>
<td>50</td>
<td>10</td>
<td>68</td>
<td>On</td>
</tr>
<tr>
<td>10:00 AM</td>
<td>28</td>
<td>52</td>
<td>9</td>
<td>70</td>
<td>On</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>29</td>
<td>54</td>
<td>8</td>
<td>72</td>
<td>Off</td>
</tr>
<tr>
<td>11:00 AM</td>
<td>30</td>
<td>55</td>
<td>7</td>
<td>73</td>
<td>Off</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>31</td>
<td>57</td>
<td>6</td>
<td>75</td>
<td>Off</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>32</td>
<td>59</td>
<td>5</td>
<td>77</td>
<td>Off</td>
</tr>
<tr>
<td>12:30 PM</td>
<td>33</td>
<td>60</td>
<td>4</td>
<td>78</td>
<td>On</td>
</tr>
<tr>
<td>1:00 PM</td>
<td>34</td>
<td>62</td>
<td>3</td>
<td>80</td>
<td>On</td>
</tr>
<tr>
<td>1:30 PM</td>
<td>35</td>
<td>64</td>
<td>2</td>
<td>82</td>
<td>On</td>
</tr>
<tr>
<td>2:00 PM</td>
<td>36</td>
<td>65</td>
<td>1</td>
<td>83</td>
<td>On</td>
</tr>
</tbody>
</table>

After training each of the ML models, the performance of the models was evaluated to determine how well they can predict water requirement for growing tomatoes. The result of the accuracy are shown in figure 2. Thus, based on the results, it is observable that the Artificial Neural Network model provided the best results, 97.6% accuracy. The high accuracy rate means that the model managed to capture the non-linear relationships in the dataset and properly predict what the irrigation requirements should have been according to the sensors. The second-best model is the Support Vector Machine, which produced 95.4% accuracy. The primary advantage of support vector machine is that it can effectively classify data of the sensors into different categories using the support vectors. Therefore, it can be utilized in the commercial application of precision agriculture to scalarize the water requirements accurately.

Figure 2. Accuracy of each model
The Decision Trees model produced an accuracy of 91.3%, and it is an effective algorithm employed for implementation, as it is based on the features generated by the sensors. Decision tree is “one of the simplest of the ML algorithms” and could significantly enhance the detection of patterns for determining the activation of the pump. Random forest has an accuracy of 88.9% and has an excellent capability of avoiding the generalization and overfitting of the data. Many decision trees are integrated with the ensemble learning approach to enhance the capabilities of the data, which show a slightly low accuracy compared to the mentioned algorithms. In conclusion, the performance of each of the model was evaluated, and it was important to analyze the performance of each ML algorithm and adjustment of parameters in the algorithm to optimize the accuracy of prediction algorithms. The results of the study would reflect the effectiveness in overcoming the challenges of water management for testing the algorithms to ensure the high productivity of crops.

The performance scores of each ML model Analytics results in figure 3, which include precision, recall, F1 score, and accuracy, can provide a detailed assessment of the efficiency of the considered ML models in predicting water requirements for the cultivation of tomatoes. Starting with Artificial Neural Network, it has a precision of 97.8%, which indicates that 97.8% of the instances which the model classified as needing irrigation, actually needed it. Moreover, the recall score of 97.4% can be observed for the rate at which the model was able to predict the instances that actually required irrigation out of all instances of the need for water. It can be noted that F1 score is 97.6%, which is calculated as the harmonic mean of precision and recall, to consider not only false positives and false negatives but also to make the assessment of the ML model more accurate. The score is also relevant to the accuracy of 97.6%, suggesting the correctness of predictions for the majority of instances. Moving to SVM, this ML model has a precision of 95.6% and a recall, as well as F1 score of 95.2%. Judges can see that all scores are slightly lower than this for ANN, which allows the conclusion that the model classifies instances into the necessary categories with similar performance. The accuracy of 95.4% can be observed as relevant to the correctness of predictors. The same situation is noted for DT that presents precision, recall, F1 score and accuracy of 91.7%, 90.9%, 91.3% and 91.3%, respectively. Finally, the last RF model has precision, recall, F1 score of 89.2%, 88.6% and 88.9%, and its accuracy is 88.9% as well. Though this result is the poorest, it can still be relevant to the assessment of water management decisions because this model generalizes well for unseen data.

The confusion matrices describe the error rate of the predictions made by each type of ML model. In the case of the Artifical Neural Network, 1020 instances were both classified correctly as negative. 1355 instances were classified correctly as positive, and only 30 instances were misclassified as negative. The ANN got confused by 15 instances which were supposed to be classified as negative but were classified as positive instead. The SVM did a somewhat better job. 1005 instances were correctly classified as negative. 1340 instances were classified correctly as positive. 45 instances were misclassified as negative and 30 examples were misclassified as positive. The DT provided 980 correct negative classifications. For positive classifications, we have 1245. 70 instances were misclassified inside the negative category. The decision tree model seems to be mistaken things for 125 also
misplaced instances in the positive category. Lastly, the RF model made 965 correct negative predictions and 1215 correct positives. Had the model been more differentiating in its behavior, we could miss 85 positive instances and 155 negative ones.

![Confusion matrices of each model](image-url)

**Figure 4. Confusion matrices of each model**

VII. CONCLUSION

The adoption of Machine Learning models into precision agriculture systems can contribute to improving the efficiency of water management procedures and consequently, increasing crop yields. By using the data collected by sensors and run through complex algorithms, including Artificial Neural Networks, Support Vector Machines, Decision Trees, and Random Forest, burning system operation can be predicted successfully, and the threat of either excessive or inadequate watering can be minimized. As far as the effectiveness of each model in the context of the research is concerned, ANN turned out to be slightly better than SVM, which, in its turn, was slightly more successful than DT and RF. The analysis of data obtained with the help of the latter has clearly demonstrated the value of applying advanced technologies to the redesign of the existing irrigation systems.

Moreover, the use of sensor networks built around IoT technologies allows for monitoring the existing environmental conditions in real-time. As a result, the danger of inappropriate reactions to the data related to the same irrigation systems can be minimized. For instance, with the help of IoT, farmers can track their water consumption even more actively and enhance their abilities to save water and, thus, significantly minimize the cost and effect that the use of different resources entailed by the manufacturing process of irrigation systems carry. As researchers make more discoveries and design new sensors for monitoring, the Internet of Things is bound to shape the future development of precision technologies.

REFERENCES


