JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Forecasting Plant Development and Production in Greenhouse Settings with Deep Learning

Andalam Srihari,
Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

V.N.S Manaswini,
Assistant Professor, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

U.Swetha,
Assistant Professor, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

V.Satheesh,
Assistant Professor, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

Abstract - This main aim of this project is to apply Deep Learning to Establish a Predictive Model for Plant Growth and Output in Greenhouse Settings. In the realm of agriculture and greenhouse cultivation, it is imperative to be able to prognosticate the growth and yield of plants. This capability affords cultivators the advantage of making the necessary adjustments to the growing environment for increased production, balancing supply and demand, and minimizing expenses. The integration of Machine Learning (ML) and Deep Learning (DL) technology presents an innovative approach to this challenge. The focus of this investigation is to use ML and DL techniques to anticipate the yield and stem growth variability of two selected crops, namely tomatoes and Ficus benjamina, in environments specifically controlled for greenhouse agriculture.

In this investigation, we advance a cutting-edge deep recurrent neural network (RNN) that utilizes the Long Short-Term Memory (LSTM) cell model for the prediction of growth. The network considers both the past records of yield, growth, and stem diameter, as well as the microclimate conditions to predict the target growth parameters. The efficiency of this approach is evaluated through comparison with other machine learning (ML) techniques such as Support Vector Regression and Random Forest Regression, using the Mean Square Error criterion. The findings of this study are obtained from data gathered from two greenhouses situated in Belgium and the United Kingdom as part of the EU Interreg SMARTGREEN project, which was conducted between 2017 and 2021, and present encouraging outcomes.

1. INTRODUCTION

The growth of plants, like many other biological systems, is intricate and subject to change due to environmental factors. The modeling of such growth and yield presents a formidable challenge to researchers. Two main approaches to this modeling have been established, as noted by Todorovski and Dzeroski (2006) and Atanasova et al. (2008), these being the "knowledge-driven" and "data-driven" approaches. The former approach is primarily founded on prior domain knowledge, whereas the latter approach derives a model only from obtained data without the necessity of relying on any pre-existing knowledge.

The category of Data Driven Models (DDMs) encompasses a range of traditional Machine Learning techniques, including artificial neural networks, support vector machines, and generalized linear models. These methods possess several attractive characteristics, such as a lack of stringent constraints, the aptitude to imitate nonlinear functions, superior predictive powers, and the adaptability to diverse inputs within a multivariate system. According to studies conducted by Singh et al. (2016) and later reviewed by Liakos et al. (2018), Machine Learning, linear polarizations, wavelet-based filtering, vegetation indices (NDVI), and regression analysis are the most commonly employed techniques for the analysis of agricultural data. Nevertheless, a relatively new methodology, known as deep learning (DL), has emerged and is rapidly gaining popularity in the realm of machine learning computations. This technique shares similarities with artificial neural networks (Goodfellow et al., 2016). As opposed to conventional neural networks, deep learning (DL) employs a series of operations to confer a hierarchical representation of data, thus promoting a greater learning capacity and a heightened level of accuracy.

A salient benefit of DL is feature learning, whereby raw data undergoes automated processing to generate features, with higher-level features being constituted from the coalescence of lower-level features (Goodfellow et al., 2016). The capability of DL to successfully address more complex issues stems from its utilization of more advanced models (Pan and Yang, 2010). Given an ample supply of data that characterizes the issue at hand, the employment of intricate models within the context of deep learning can bring forth an augmentation in the precision of classification endeavors and a diminution of error in regression challenges.

MATERIALS AND METHODS

Machine learning, as a computational discipline, affords the potential to tackle complex, non-linear problems via the utilization of data derived from a multitude of sources. This methodology permits the realization of informed decision-making processes with limited human intervention, thereby constituting a widely applied framework for the purpose of data-driven decision making, particularly in the realm of agriculture. In recent times, various techniques of machine learning have been employed with a view to foretelling plant growth, harvest, and production for crops of varying varieties. Such techniques encompass Artificial Neural Networks, Support Vector Regression, M5-prime Regression Trees, Random Forests, and K-Nearest Neighbors, which have been demonstrated to effectively enhance the accuracy of predictions related to plant growth this examination, the Support Vector Regression (SVR) and Random Forests (RF) models are taken as the standard models for forecasting plant growth and yield. SVR originates from a non-linear version of the Generalized Portrait formula created by Vapnik. It functions by transforming the input data into a higher dimensional space through the utilization of a kernel operation, and afterwards dividing the data into separate groups with a hyperplane.

The extent of error and margin is managed by a regulation parameter c. The SVR with radial basis kernel procedure (SVRrbf) employs a formula (K) that assesses the similarity between two data points based on their proximity. The formula is expressed as $(xi,xj) = \exp(-y\|xi - xj\|2)$. In this study, the Random Forest (RF) approach is also taken into consideration as a benchmark for determining plant growth and yield. RF is a category of ensemble learning algorithms, which was originally proposed

by Ho in the year 1998. The methodology of RF is built upon the principle of decision trees as its primary learner, since it is believed that a sole predictor is not capable of accurately predicting the desired value of test data, as it fails to distinguish between random noise and meaningful patterns in sample data. To overcome this limitation, RF creates numerous autonomous regression trees by selecting a subset of the training data, referred to as a bootstrap sample, at each tree. These regression trees are grown until they reach the maximum possible size, and the final prediction is obtained through taking a weighted average of the predictions made by all the regression trees (as described by Breiman in 2001)

2. LITERATURE SURVEY

Deep Learning represents an advancement in the realm of Machine Learning, with the introduction of greater intricacy into the model and the transfiguration of data through multiple levels of abstraction in a hierarchical manner. This technique boasts the advantage of feature learning, wherein the extraction of features from raw data is performed automatically and higher-level features are produced through a combination of lower-level features. Owing to the presence of more elaborate models in DL, it has the ability to tackle complex problems with speed and efficiency, and the models can be massively parallelized for increased performance. Enhanced classification accuracy or reduced regression error can be achieved through the use of DL, given the availability of a sufficiently large dataset describing the problem. The components included in DL vary based on the network architecture utilized, such as Convolutional Neural Networks, Recurrent Neural Networks, and Unsupervised Networks (Kamilaris et al., 2018).

The Long short-term memory (LSTM) model was first introduced by (Hochreiter and Schmidhuber, 1997) for the purpose of modeling long-term dependencies and determining the optimal time lag in time series problems. The architecture of the LSTM network comprises of an input layer, a recurrent hidden layer, and an output layer. The hidden layer contains memory blocks that incorporate memory cells that retain temporal state information and two adaptive, multiplicative gating units that control the flow of information within the block. The memory cell essentially operates as a recurrent linear unit, known as the Constant Error Carousel, and its state is depicted by the activation of the CEC. The multiplicative gates learn when to open and close, and the network error is kept constant, resolving the issue of vanishing gradient in LSTM. Furthermore, a forget gate has been added to the memory cell to prevent the gradient from becoming unstable when learning long time series.

Two distinct models were created to investigate the correlation between environmental conditions in greenhouses and tomato yield. The first model, established by Abreu et al. (2000), analyzed the correlation between fruit growth and flowering rate in greenhouses located in the southern region of France that utilized heating. However, when the model was applied in plastic greenhouses located in Portugal that did not use heating, its accuracy was found to be insufficient. The second model, put forth by Adams (2002), aimed to demonstrate weekly alterations in tomato yield in greenhouses through the representation of fruit size and harvest rate. The hourly climate data was employed to determine the rate of growth for leaf truss and flower production. The yield's periodic fluctuations were determined to be impacted by changes in solar radiation and air temperature. Qaddoum et al. (2013) stated that there exist numerous tools available to aid farmers in decision-making, which provide yield rate predictions, propose climate control techniques, and align crop production with market demands.

3. SYSTEM ANALYSIS:

The purpose of incorporating the Long Short-Term Memory (LSTM) model is to assess and identify long-range dependencies and determine the most appropriate time delay for time-series analysis. The composition of an LSTM network includes an input layer, a recurrent hidden layer, and an output layer. The central component within the hidden layer is the memory block, which is comprised of memory cells

that maintain the temporal state and two adjustable, multiplicative gate units that control the flow of information in the block. The memory cell primarily functions as a recurrent self-connected linear unit referred to as the Constant Error Carousel (CEC), and the state of the cell is indicated by the activation of the CEC. The multiplicative gates learn when to be opened and closed, and by keeping the error constant within the network, the vanishing gradient problem can be effectively addressed in LSTM. Furthermore, a forget gate is added to the memory cell to prevent the gradient from increasing excessively when analyzing lengthy time-series.

This project comprises of the following phases:

- Procuring the Ficus plant dataset.
- Purifying the dataset by discovering and exchanging any absent information with either the average or 0.
- Partitioning the dataset into training and testing segments, where 80% is used to train the machine learning algorithms and 20% is employed to assess the accuracy of predictions through Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).
- Executing the Support Vector Regression (SVR) and determining its proficiency with the test data.
- Executing the Random Forest (RF) and determining its proficiency with the test data.
- Executing the Long Short-Term Memory (LSTM) and determining its proficiency with the test data.
- Utilizing the LSTM classifier to anticipate plant growth and yield based on the test data.

This project involves the utilization of the Ficus plant dataset, which is stored within the designated 'dataset' folder. The following are a few examples of the dataset.

4. SYSTEM DESIGN:

4.1 SYSTEM ARCHITECTURE

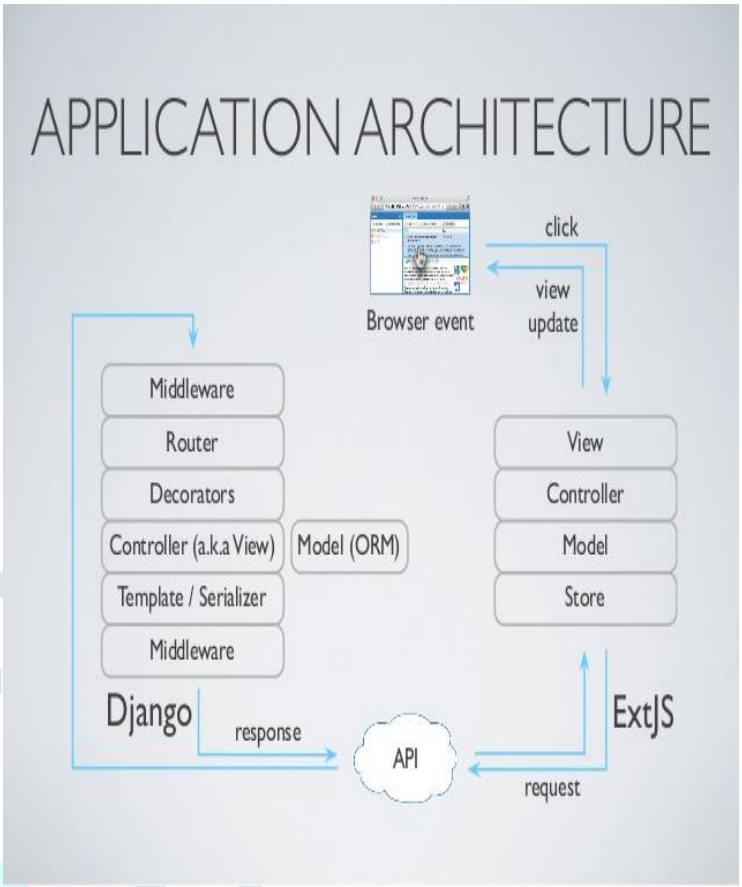
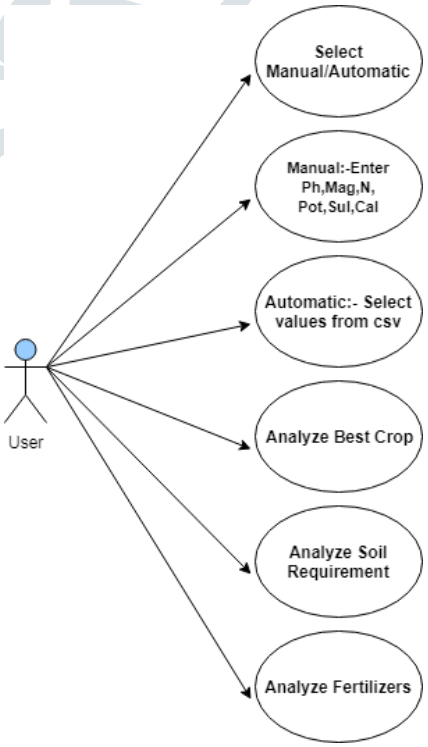
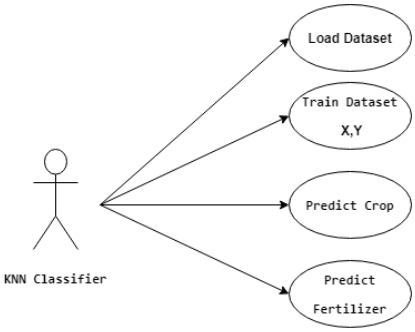


Figure 4.1: Architecture diagram

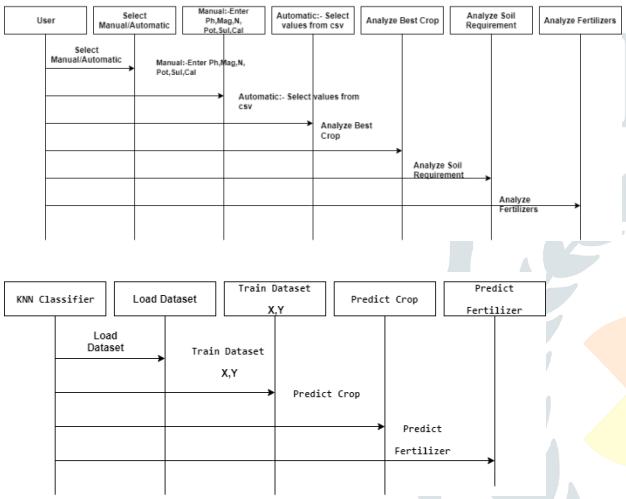
4.2 UML DIAGRAMS

4.2.1 USE CASE DIAGRAM

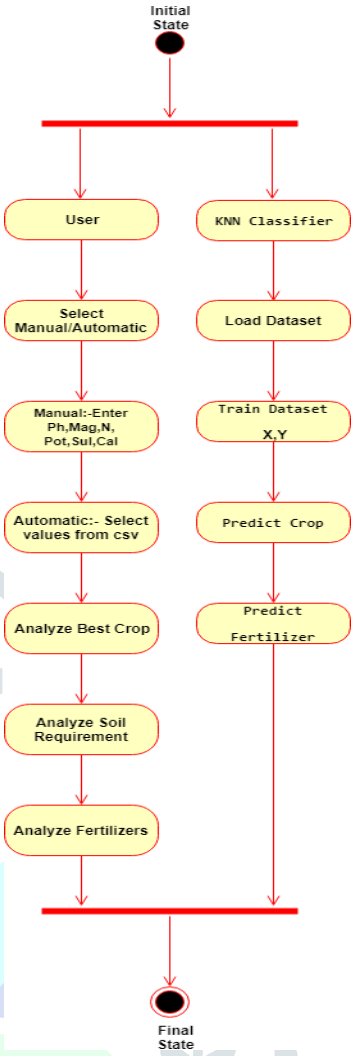




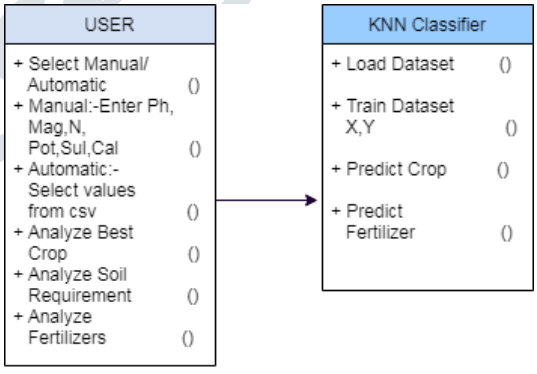
4.2.3 SEQUENCEDIAGRAM



4.2.4 ACTIVITY DIAGRAM



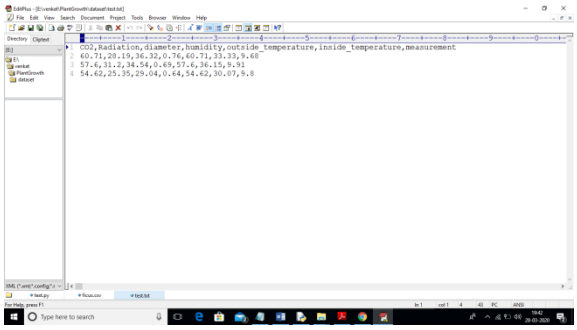
4.2.5: Class Diagram

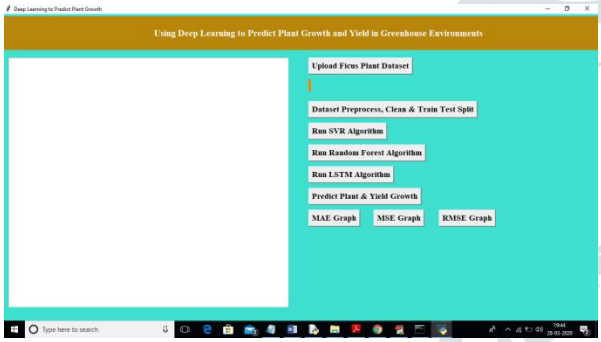


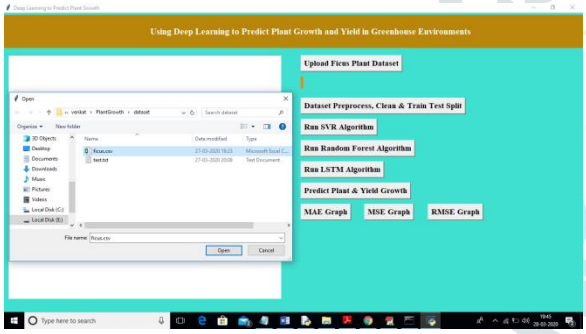
5. SCREEN SHOTS

In this project, the Ficus plant dataset, located in the "dataset" folder, is employed. The dataset encompasses columns like CO2, RADIATION, DIAMETER, etc. with the last column indicating the yield of the crop under the specified environmental factors. By training a classifier with these values, we

will be able to input test data and predict future growth or yield. The test data comprises of environmental factors, with the YIELD column absent, and the classifier will determine this missing value.







The given data set consists of environmental values, but the growth or yield values are missing. By using the LSTM classifier on this data, future growth predictions can be made. To proceed, one must double-click on the "run.bat" file and then click on the "Upload Ficus Plant Dataset" button to upload the "ficus.csv" dataset. The application will then split the dataset into 80% for training and 20% for testing, using 3222 records for training and 806 for testing. Next, the SVR and Random Forest algorithms will be trained on the dataset, followed by the LSTM algorithm. The LSTM algorithm proves to have a lower mean squared error, root mean squared error, and mean absolute error compared to the other algorithms. Finally, by uploading a "test.txt" file, one can predict the growth of new records and view a comparison graph between all algorithms. The graph shows that the LSTM algorithm has a lower error and provides the best predictions. To generate the LSTM training model, the application uses 10 epochs, where in each epoch the LSTM uses recent data to train the model while forgetting the old references.

6. CONCLUSION

The article details a Deep Learning (DL) method utilizing Long Short-Term Memory (LSTM) for predicting both the stem diameter variation (SDV) of the Ficus plant and tomato yield. The findings reveal that this DL technique, utilizing an LSTM model, surpasses other traditional Machine Learning (ML) methods such as Support Vector Regression (SVR) and Random Forest (RF) in terms of accuracy as measured by mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) metrics. Consequently, the purpose of our initiative is to devise Deep Learning (DL) approaches for forecasting plant growth and yield in greenhouses under controlled conditions. Prospective pursuits aim at enhancing the DL methods by increasing the amount of data utilized for training and expanding the methods for multi-step forecasting, with a weekly or multiple-week interval, across various greenhouses in the United Kingdom and Europe.

REFERENCES

- Abreu, P., Meneses, J. & Gary, C. 1998, "Tompousse, a model of yield prediction for tomato crops
- Electronic Information in Horticulture 519, pp. 141.
- Cortes, C. & Vapnik, V. 1995, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp. 273-297.
- Heuvelink, E. 1996, Tomato growth and yield: quantitative analysis and synthesis, Heuvelink.