Chapter 15
LSTM-Based Deep Learning for Crop Production Prediction With Synthetic Data

Aditi Verma
School of Computer Science Engineering, Vellore Institute of Technology, India

Shivani Boggavarapu
http://orcid.org/0009-0003-2094-1449
School of Computer Science Engineering, Vellore Institute of Technology, India

Astha Bharadwaj
School of Computer Science Engineering, Vellore Institute of Technology, India

Prabakaran N.
http://orcid.org/0000-0002-1232-1878
School of Computer Science Engineering, Vellore Institute of Technology, India

ABSTRACT

The Agri-industry forms the backbone of the economy and livelihood. Hence, efficient planning on resources and ensuring a steady food supply is vital. This model discusses the challenges of accurately predicting crop yields influenced by multiple dynamic factors. Traditional models suffer with the complexity, thus leading to inaccurate predictions. Also, the availability of reliable training data is scarce, which poses an additional problem in training. Existing solutions range from traditional statistical models based on historical data to modern AI techniques. While these approaches are better than conventional methods, they are still unable to address data scarcity, non-linear interactions and the dynamic complexities. This model aims to overcome the limitations using long short-term memory (LSTM) and integrating synthetic data. LSTM is able to decipher complex patterns and synthetic data provides additional training samples that can enhance accuracy. The overall potential of this proposed solution can help mitigate food scarcity and strengthen sustainability.

DOI: 10.4018/979-8-3693-3583-3.ch015

INTRODUCTION

Modern agriculture faces the challenge of ensuring food security for a growing global population while dealing with unpredictable climate patterns and limited resources. Predicting crop production accurately is crucial for efficient resource allocation, risk management, and policy planning. Deep Learning techniques have emerged as powerful tools for time series prediction tasks due to their ability to capture complex temporal dependencies. Furthermore, the integration of synthetic data augments the training process, enhancing model performance and generalization (Prodhan, F. Et al., 2022). The fusion of LSTM networks and the integration of synthetic data offer a promising solution for accurate crop production prediction. This approach not only addresses the limitations of traditional methods but also enables proactive decision-making in agriculture, contributing to global food security and sustainable resource management. Utilizing data-driven methods has become crucial in modern agriculture for improving crop output prediction. The Long Short-Term Memory (LSTM) method, a kind of recurrent neural network (RNN), is a potent tool in this attempt. Due to its design for modeling sequential data, LSTMs are well suited for the time-dependent and dynamic character of agricultural production datasets. It particularly adept in capturing the complex temporal relationships and patterns present in the variables influencing agricultural yields, such as weather patterns, soil moisture, and past output rates. LSTMs, as opposed to conventional RNNs, are able to successfully learn from both short-term and long-term information because they are able to solve the vanishing gradient problem. By consuming a series of input data representing the changing conditions over time, the LSTM can be used to predict future crop yields once it has been trained. The latter half of this approach incorporates synthetic data generation and integrating it to the training model. Current approaches of crop yield prediction (Joshua et al., 2022) which involve machine learning algorithms like random forests and support vector machines, often lack enough training samples for the model prediction. This leads to inept accuracies in the prediction results. To tackle this problem, we generate synthetic data and integrate it to the already existing real crop data set, to increase the number of training samples, thereby, increasing accuracy of the prediction model (Elavarasan, D., & Durai Raj Vincent, 2021) and (Van Klompenburg et al., 2020). Synthetic data is the artificially generated data that replicates or mimics the real-world data. It does not contain any personally identifiable information (PII). As mentioned previously, it is used for a variety of purposes, including data analysis, model training and testing. Synthetic data can be generated using algorithms or statistical models based on the characteristics and patterns observed in real data. Some of the ways of generating synthetic data are rule-based generation, simulation models, generative adversarial networks (GANs), variational autoencoders (VAEs) etc. The algorithm which we will be deploying to generate synthetic data for our dataset is the Tabular LSTM model. It is a type of LSTM model which can specifically handle tabular data (structured into rows and columns). Tabular LSTMs use the sequential and temporal aspects of LSTM architecture to process tabular data with sequential dependencies, making them favorable for time-series prediction. The reason we chose LSTM for both synthetic data generation and crop yield prediction is because LSTM models are most suitable for structured and sequential data. Crop data collected for crop prediction consists of attributes such as weather and soil parameters, which can be organized in a structured format. Furthermore, this structured data becomes sequential when the temporal order matters, as the crop growth is influenced by the progression of time and changing conditions.

Related Content