**Datasets to Use:**   
Worldwide - Kaggle:  
<https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset>  
USA Corn- Syngenta Crop Challenge:   
<https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php/>  
Illinois, Indiana, Iowa:

<https://github.com/ansarifar/An-Explainable-Model-for-Crop-Yield-Prediction>

**Papers to Read:**   
2017: Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data  
<https://cs.stanford.edu/~ermon/papers/cropyield_AAAI17.pdf>

2019: Crop Yield Prediction Using Deep Neural Networks

<https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.00621/full>

2020: Crop yield prediction using machine learning: A systematic literature review

<https://www.sciencedirect.com/science/article/pii/S0168169920302301>

* 2024: Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability

<https://www.sciencedirect.com/science/article/pii/S2405844024168673#fig3>

* 2025: Crop yield prediction using machine learning: An extensive and systematic literature review

<https://www.sciencedirect.com/science/article/pii/S2772375524003228>

2021: An interaction regression model for crop yield prediction  
<https://www.nature.com/articles/s41598-021-97221-7>

2023; A data-driven crop model for maize yield prediction

<https://www.nature.com/articles/s42003-023-04833-y>

**Crop Yield Production Algorithms:**  
Decision Tree: <https://github.com/Chando0185/Crop_Yield_Prediction>

Deep Learning: <https://github.com/JiaxuanYou/crop_yield_prediction>

CNN RNN Framework: <https://github.com/saeedkhaki92/CNN-RNN-Yield-Prediction>

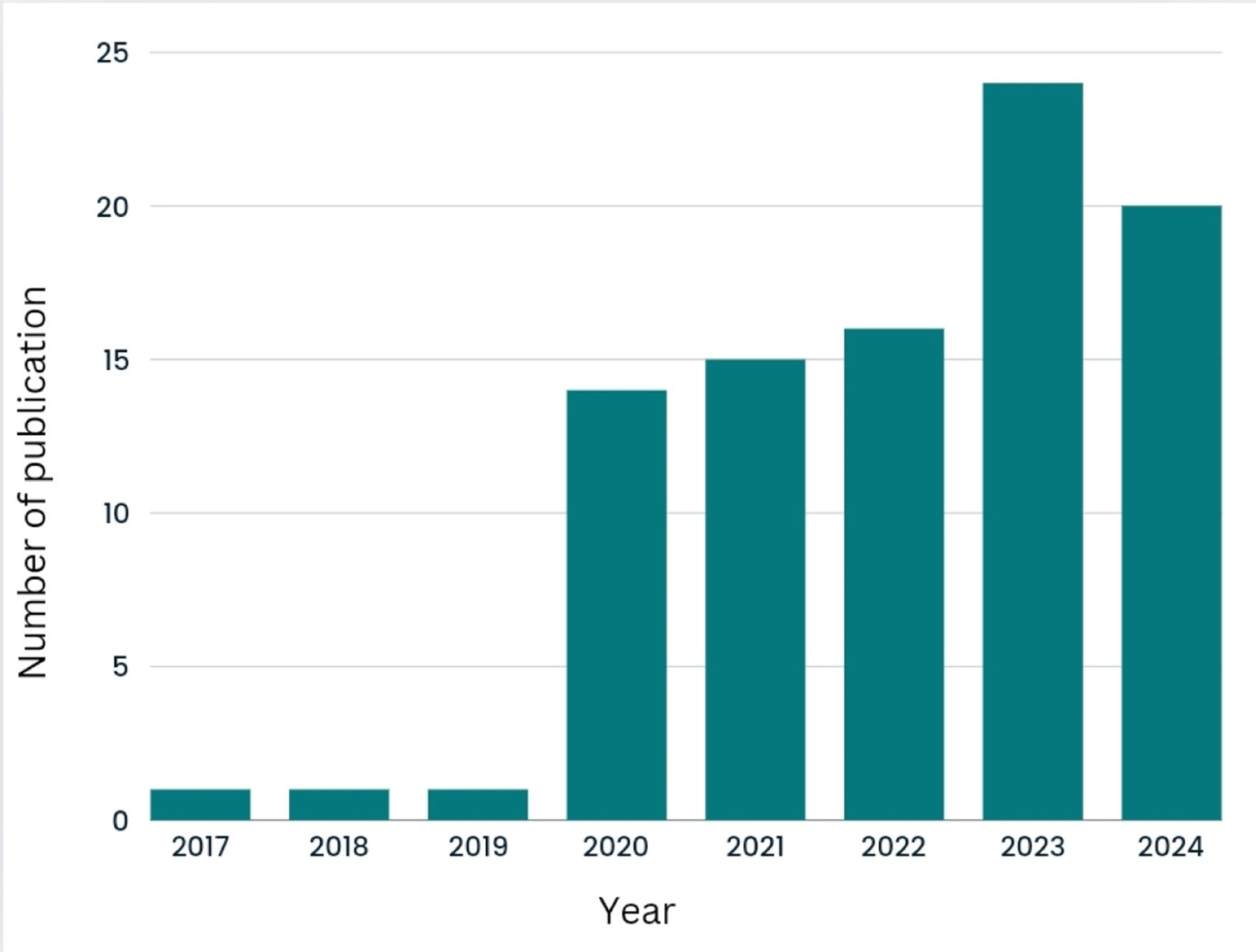
Deep Gaussian Process: <https://github.com/gabrieltseng/pycrop-yield-prediction>

**Cool Videos:**  
<https://www.youtube.com/watch?v=LEnjAR6AwOo>

# **Crop Yield Prediction Notes**

Overarching facts:

* WHO estimate: 820 million people worldwide still have insufficient access to food
* Food and Agriculture Organization (FAO) has projected a 60 percent increase in the demand for food to meet the needs of the projected global population of 9.3 billion by 2050

97 Total Articles regarding this topic

## **Purpose**

* Utilize models to **optimize the allocation of resources** such as water, fertilizer, and pesticides, thereby maximizing yields and minimizing input costs
* By anticipating potential yield fluctuations due to environmental factors, farmers can **implement risk management strategies** to mitigate the impact of adverse weather conditions or other environmental stressor
* Enable stakeholders to make informed decisions regarding market planning, trade policies, and commodity pricing, contributing to more efficient and resilient agricultural markets
* Develop **evidence-based policies** and programs aimed at promoting food security, sustainable agriculture, and rural development
* Underlying crop yield prediction is also a fundamental research question in plant biology, which is to understand how plant phenotype is determined by genotype (G), environment (E), management (M), and their interactions.

## **Guiding Questions**

1. What are the applied machine learning algorithms for crop yield prediction?
2. Which features were used for experiments in previous studies?
3. What evaluation metrics are used for model performance evaluation?
4. What difficulties were faced by the researchers while conducting research?

## 

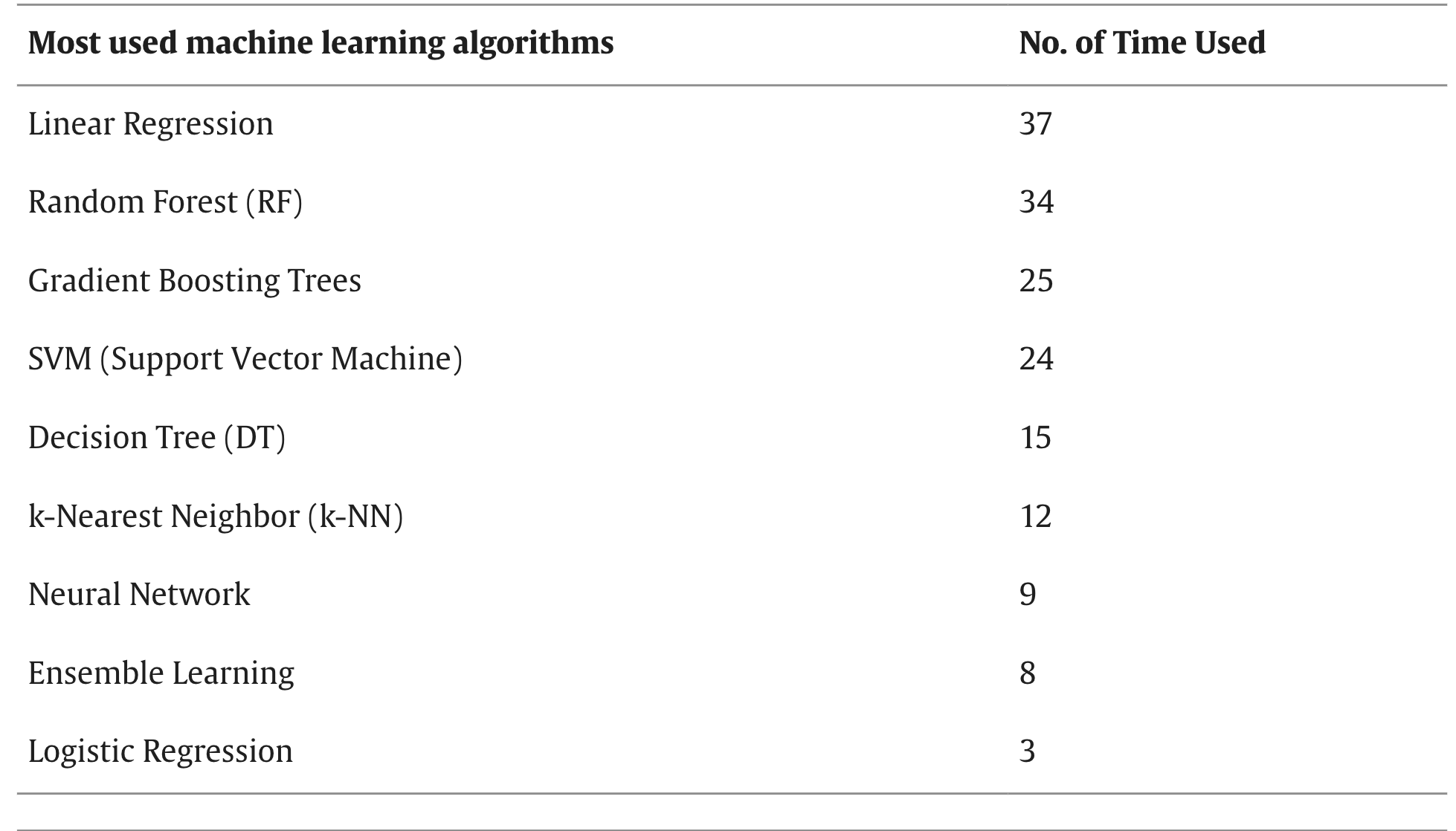
## 

## 

## 

## 

## **1a. Most used machine learning algorithms**

  
Agricultural data often contain noisy features and outliers, which regularization can help mitigate

Linear regression often served as comparison tool

### **Support Vector Machine (SVM)**

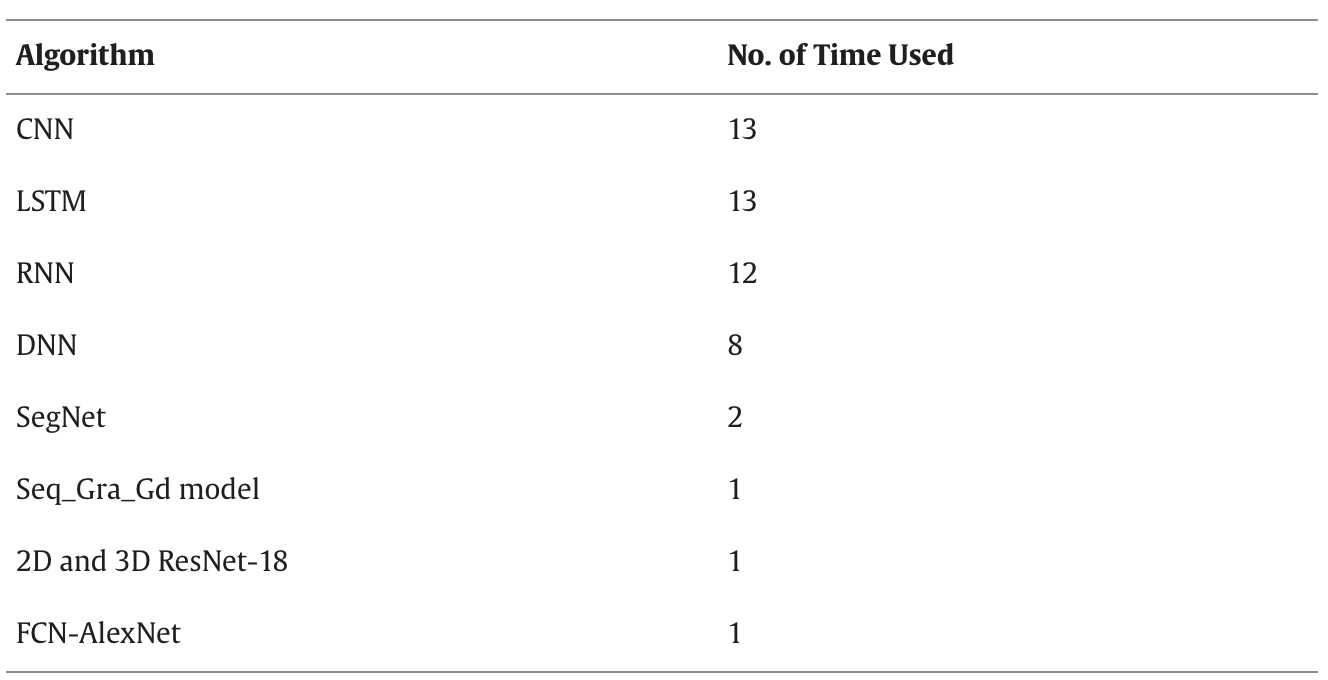
Regularized classifiers like linear SVM outperform non-regularized methods like LDA, and SVM's decision rule is a reliable and low-variance linear function in kernel space

* SVM performs well with small training sets and high-dimensional feature vectors
* Can classify high dimensional feature vecotr

### **Random Forest (RF)**

* offers **high-speed operation, generalization performance, and an ensemble of tree-structured classifiers**, making it ideal for finding non-linear patterns
* requires minimal data preprocessing and excels in feature selection for superior yield results
* Can manage high dimensional dataset

## **1b. Most used Deep learning algorithms**



### **Recurrent Neural Network (RNN):**

An RNN's hidden state is determined by the inputs at the current time step as well as the hidden state from the time step prior. This serves as the network's "memory," storing data about the content that has already been viewed. The RNN updates its hidden state for every input in the sequence, which enables it to extract context from the sequence's past. Unlike other neural network types, an RNN can process sequences of variable lengths as a result of this feature. RNN has shown a low error rate for crop yield prediction.

### **Convolutional Neural Network (CNN):**

CNNs are widely used in various applications due to their flexibility and ability to automatically identify hierarchical features from visual input. A typical CNN model consists of several convolutional and pooling layers, as well as a small number of fully connected (FC) layers. The design of CNNs involves several parameters, such as the number of filters, filter size, type of padding, and stride [51]. There are two methods to implement an architecture: using pre-defined architectures or starting from scratch. The first method's drawback is that it requires a lot of time, but the second approach offers time savings and superior performance.

### **Deep Neural Network (DNN):**

A DNN is made up of follow-up layers of neurons that lead to an output layer. These layers can be viewed as sequential representations of the data being entered. The DNN architecture uses neurons as elementary computing units with an activation function to model layers connected using weighted links. During the training phase, the architecture evaluates weight vectors and uses forward and backward passes to compute prediction errors and gradients.

### **Long short-term memory (LSTM):**

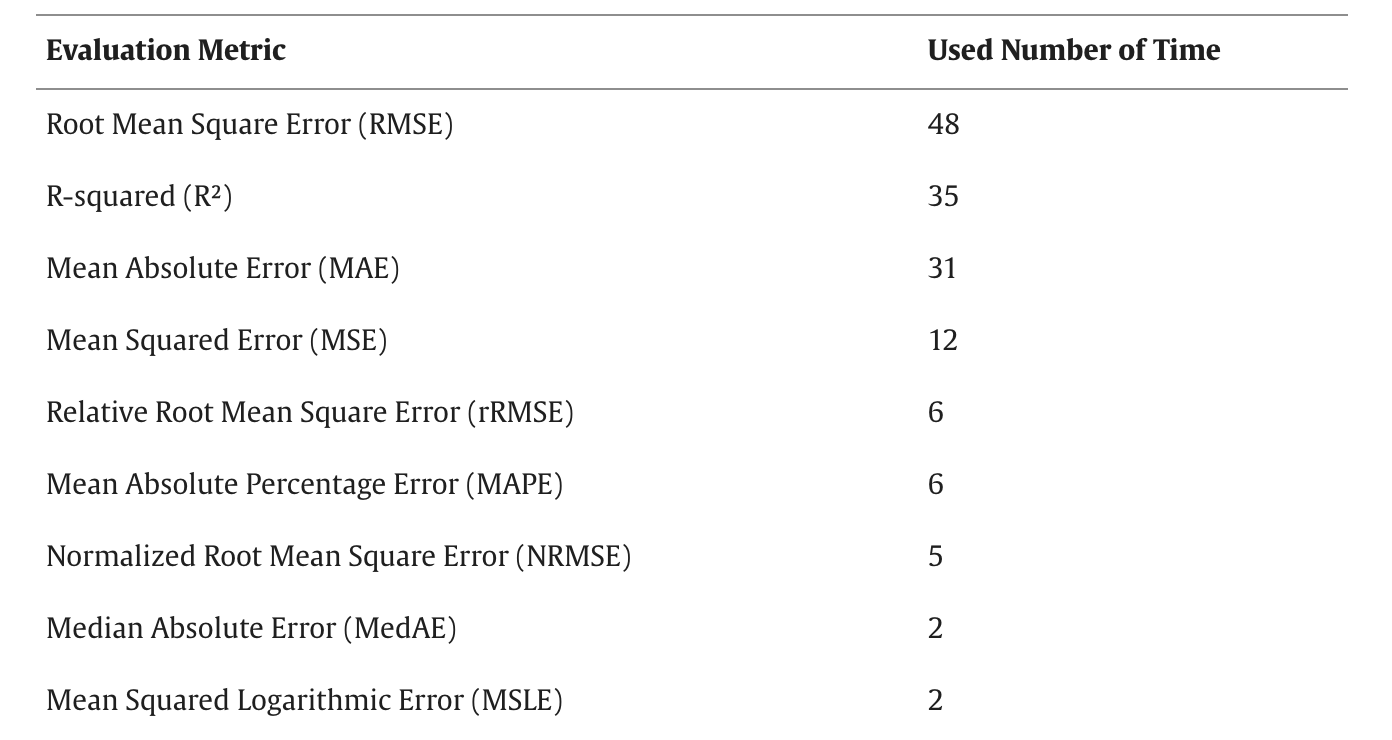
Recurrent neural networks (RNNs) frequently utilize the Long Short-Term Memory (LSTM) architecture for sequence modeling applications. The LSTM architecture is characterized by a more intricate structure that includes extra memory cells and gates, enabling it to selectively retain or discard information from preceding time steps. By taking into account past historical values, LSTM examines current unknown patterns and adapts itself based on the complete patterns, enabling it to make future predictions [13].

\*\*\*Add on Hybrid and State-of-the-art Models  
\*\*\*Deep learning Algorithms + machine learning methods like decision trees, random forests, and XGBoost regression

## **2. Most used Features**

For example, all soil variables such as soil type, soil pH, and soil texture were grouped into Soil Information. Crop Information included variables like crop types, crop weight, and crop growth stages. Weather Information grouped weather data such as temperature, rainfall, and precipitation. Field Management included field adjustment variables like irrigation data, plant spatial arrangement, and field sizes. Humidity grouped variables such as humidity, vapor, and vapor pressure. All types of image data were categorized under Imagery Data. Lastly, features that did not fit into any of the predefined categories were compiled into a distinct category labeled 'Other.'

## **3. Most used Metrics**



## **4. Problems**

* Insufficient Datasets
  + variations in climatic conditions
  + different types of vegetation
  + longer time series of yield data
* implementing the models in agricultural systems