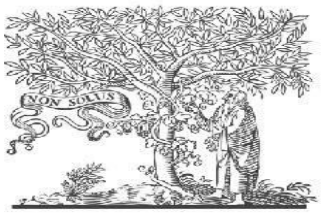
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

Link : https://ijiemr.org/downloads/Volume-12/Issue-05

10.48047/IJIEMR/V12/ISSUE05/19

Title **Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers**

Volume 12, Issue 05, Pages: 179-189

Paper Authors

J. Ravichandra Reddy, . G. Vinay, K. Ramya, M. Namratha, P. Sanjay Kumar

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers

1. **J. Ravichandra Reddy** ,. Assistant Professor, Mtech (Phd), Department of CSE, (Vijay Rural Engineering College(VREC)) ravichand48@gmail.com

2. **G. Vinay,** BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) gampa.vinnu2024@gmail.com

3. **K. Ramya**, BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) ramya01122000@gmail.com

4. **M. Namratha**, BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) namrathamegharaj99@gmail.com

5. **P. Sanjay Kumar,** BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) sanjaysmarty173@gmail.com

ABSTRACT: Agriculture is a developing topic of study. Crop prediction, in particular, is crucial in agriculture and is heavily reliant on soil and environmental factors such as rainfall, humidity, and temperature. Farmers used to be able to choose the crop to produce, watch its progress, and select when it might be harvested. Today, however, fast changes in environmental circumstances have made it impossible for farmers to continue in this manner. As a result, machine learning approaches have taken up the role of prediction in recent years, and this study has employed many of them to calculate crop production. To guarantee that a particular machine learning (ML) model operates with high accuracy, it is critical to use effective feature selection techniques to preprocess raw data into a Machine Learning friendly dataset. Only data characteristics that have a high degree of importance in defining the final output of the model must be used to eliminate redundancy and improve the accuracy of the ML model. As a result, optimum feature selection emerges to guarantee that only the most relevant characteristics are included in the model. Consolidating every single characteristic from raw data without considering their importance in the model-making process would unduly complicate our model. Furthermore, adding characteristics that contribute little to the ML model would raise its time and space complexity, affecting the model's output accuracy. The findings show that an ensemble

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

method outperforms the current classification technique in terms of prediction accuracy.

Keywords *– Agriculture, classification, crop prediction, feature selection.*

1. INTRODUCTION

Crop prediction in agriculture is a complex process, with several models suggested and tested to that aim. Given that crop cultivation is dependent on both biotic and abiotic variables, the challenge necessitates the use of a variety of datasets. Biotic factors are environmental components that develop as a consequence of the direct or indirect action of living species (microorganisms, plants, animals, parasites, predators, pests) on other living organisms. Anthropogenic variables are also included in this category (fertilization, plant protection, irrigation, air pollution, water pollution and soils, etc.). These issues may cause various variations in crop production, such as internal faults, form defects, and changes in the chemical makeup of the plant output. Abiotic and biotic elements impact the formation of the environment as well as the development and quality of plants. Abiotic factors are classified as physical, chemical, or other. Mechanical vibrations (vibration, noise), radiation (e.g., ionising, electromagnetic, ultraviolet, infrared); climatic conditions (atmospheric pressure, temperature, humidity, air movements, sunlight); soil type, topography, soil rockiness, atmosphere, and water chemistry, particularly salinity, are among the recognised physical factors. Priority environmental poisons such as sulphur dioxide and derivatives, PAHs, nitrogen oxides and derivatives, fluorine and its compounds, lead and its compounds, cadmium and its compounds, nitrogen fertilisers, pesticides, and carbon monoxide are among the chemical components. Mercury, arsenic, dioxins and furans, asbestos, and aflatoxins are among the others. Abiotic elements such as bedrock, relief, climate, and water conditions all have an impact on its qualities. Soil-forming variables have a wide range of effects on soil formation and agricultural value.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

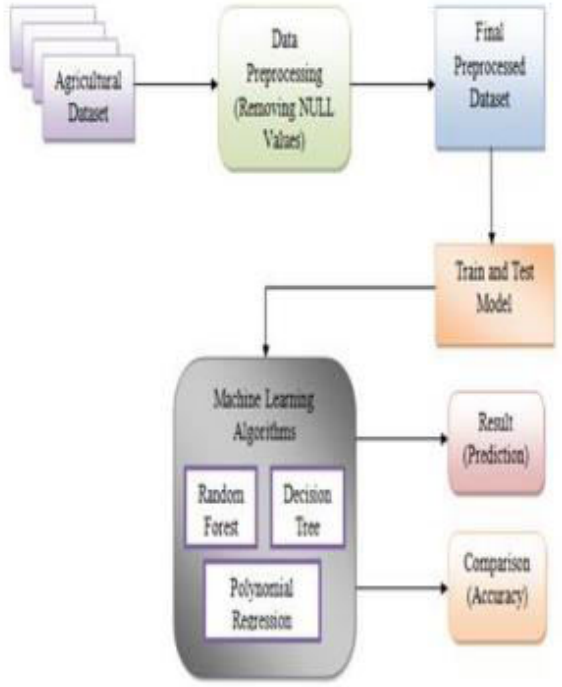


Fig.1: Example figure

Crop production prediction is neither straightforward nor easy. According to Myers et al. [5] and Muriithi [6,] the approach for forecasting the area under cultivation is a collection of statistical and mathematical strategies helpful in an ongoing and improving optimization process. It is also useful in the design, development, and formulation of new and improved goods. The presentation or conduct of statistical analysis necessitates the availability of numerical data. Based on them, inferences about diverse occurrences are derived, and binding economic choices may be taken. According to Muriithi [6,] the more you represent particular occurrences numerically, the more you can say about them, and improving data quality allows you to receive more precise information and make more correct judgements.

2. LITERATURE REVIEW

Applying naive Bayes classification technique for classification of improved agricultural land soils:

The advancements in computers and data storage have resulted in massive volumes of data. The problem has been to extract information from this raw data, which has led to the development of new approaches and techniques such as data mining, which may bridge the knowledge gap. The goal of this study was to evaluate these novel data mining methods and apply them to a soil science database to see whether relevant associations could be discovered. The Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati, has provided a big data collection of soil databases. The database comprises soil profile measurements from numerous places in Chandragiri Mandal, Chittoor District. The study determines whether or not soils are classified using different data mining approaches. Furthermore, a comparison of Naive Bayes classification and analysis of the most successful approach was performed. The study's findings might have a wide range of applications in

agriculture, soil management, and environmental protection.

Biotic components influencing the yield and quality of potato tubers

For the previous decade, potato yields in Canterbury have remained stable at about 60 t/ha. Potato growth models, on the other hand, anticipate potential yields of up to 90 t/ha, which have already been attained by certain commercial producers. A two-year study led by industry and academic partners investigated agricultural production constraints. During the first growing season, 11 processing crops were thoroughly evaluated (final yield, plant health, and soil quality tests). Soil-borne illnesses (Rhizoctonia stem canker and Spongospora root infection), as well as subsurface soil compaction and insufficient irrigation management, were found as persistent causes in lower yields. Cropping histories that contained potatoes within the past ten years led in a quicker development of Rhizoctonia stem canker symptoms (by emergence) when compared to areas with periods of grass growth and no prior potato crops (8 weeks after emergence). A controlled field experiment in a commercial crop (known to have high levels of soil-borne pathogens) was conducted in year 2 to identify and quantify the effects of soil-borne illnesses on yield. Soil fumigant (90, 112 and 146 kg/ha chloropicrin), in-furrow application of azoxystrobin (1.5 l/ha) or flusulphamide (400 ml/ha), and no pesticide control were among the treatments. Soil-borne pathogen DNA assays before and after treatment revealed a small decrease in Rhizoctonia solani and Spongospora subterranea DNA levels in the soil (plots treated with fumigant), but the findings were very varied. The final total fresh yield averaged 58 t/ha and was unaffected by treatment. The severity of R. solani on subterranean stems was continuously lower for the azoxystrobin treatment compared to all other treatments throughout the season.

Response surface methodology: A retrospective and literature survey

RSM is a combination of statistical design and numerical optimization methods used to improve processes and product designs. The initial research in this field goes back to the 1950s and has been extensively employed, particularly in the chemical and process industries. RSM has experienced extensive application and several innovative improvements during the past 15 years. This overview focuses on RSM activity since 1989. We review current study fields and suggest some subjects for future research.

Application of response surface methodology for optimization of potato tuber yield

The author studies the operational parameters necessary for maximum potato tuber yield production in Kenya. This will assist potato producers in avoiding additional input costs. The potato manufacturing process was improved using factorial design 2 3 and response surface approach. Using response surface approach, the combined impacts of water, nitrogen, and phosphorus mineral nutrients were explored and optimised. The best production parameters for potato tuber yield were determined to be 70.04% irrigation water, 124.75kg/ha of nitrogen provided as urea, and 191.04kg/ha of phosphorus supplied as triple super phosphate. A potato tuber yield of 19.36Kg/plot of 1.8meters by 2.25meters may be obtained under ideal conditions. Increased potato production may help smallholder potato farmers in Kenya better their livelihoods and save them money on inputs. Finally, I hope that the methodology used in this potato study may be extended to other commodities studies, resulting in a greater knowledge of total crop productivity.

Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning

Accurate high-resolution yield maps are critical in precision agriculture for detecting geographic yield variability patterns, pinpointing important variables driving yield variability, and offering site-specific management insights. Using remote sensing technology, cultivar variations may have a considerable impact on potato (Solanum tuberosum L.) tuber production forecast. The goal of this research was to enhance potato production prediction by combining cultivar information with machine learning approaches employing unmanned aerial vehicle (UAV) remote sensing. In 2018 and 2019, small plot tests with various cultivars and nitrogen (N) rates were carried out. Throughout the growth season, UAV-based multi-spectral photos were gathered. To incorporate multiple vegetative indicators with cultivar information, machine learning methods such as random forest regression (RFR) and support vector regression (SVR) were utilised. It was discovered that spectral data from UAVs collected during the early growing season at the tuber start stage (late June) were better connected with potato marketable yield than spectral data collected during the later growing season at the tuber maturity stage. However, the highest performing vegetative indicators and the greatest time for predicting potato output differed per cultivar. The RFR and SVR models performed poorly when just remote sensing data were used (R2 = 0.48-0.51 for validation), but greatly improved when cultivar information was included (R2 = 0.75-0.79 for validation). It is found that utilising machine learning techniques to

combine high spatial-resolution UAV photos and cultivar information may greatly enhance potato production prediction over approaches that do not use cultivar information. More research is required to enhance potato production prediction utilising more specific cultivar data, soil and landscape characteristics, management data, and sophisticated machine learning algorithms.

3. METHODOLOGY

The most difficult difficulty in the temperate temperature zone is assessing agroclimatic parameters that influence the production of winter plant species, primarily grains. The main factor impacting wintering yield, which offers access to days with temperatures above 5 degrees Celsius, their quantity and frequency, and the number of days in the wintering period with temperatures above 0 degrees Celsius and 5 degrees Celsius. A number of them may be approximated using public data and result in regression statistics in years. Models for assessing the scenario that evaluate whether they wish to be a trial of state policy in the area of grain market intervention have been developed. Predictions of agrometeorological parameters is required for accurate production forecasting. Aspects of these components' fluctuation may constitute a special challenge. Many researchers have attempted, with varied degrees of success, to address this problem.

Disadvantages:

1. Crop prediction in agriculture, in particular, is crucial and is heavily reliant on soil and environmental factors such as rainfall, humidity, and temperature.
2. Rapid changes in environmental circumstances have made it impossible for farmers to continue farming.

This research field faces a variety of problems. Crop prediction models now provide reasonable results, but they might do better. This research proposes an improved crop forecast model that overcomes these difficulties. The prediction procedure is based on two essential techniques: feature selection [FS] and classification. Prior to using FS methods, sampling approaches are used to balance an unbalanced dataset.

Advantages:

1. To avoid redundancies and improve the accuracy of the ML model, only data characteristics that have a high degree of importance in deciding the model's final output should be included.
2. An ensemble method outperforms the previous classification technique in terms of prediction accuracy.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

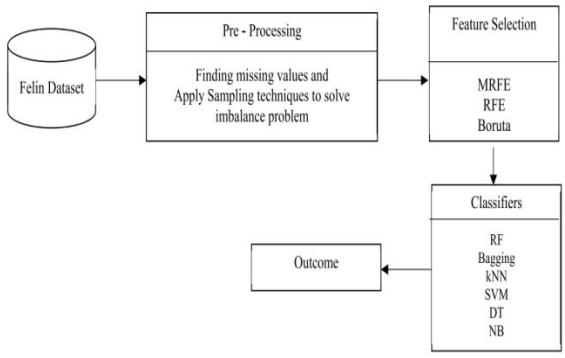


Fig.2: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- Data exploration: we will put data into the system using this module.
- Processing: we will read data for processing using this module.
- Using this module, data will be separated into train and test.
- Model generation: Building the model with and without feature selection. - Feature Choice (SMOTE, ROSE, RFE, MRFE, BORUTA, MEMOTE) - Naive Bayes - KNN - Bagging Classifier - Random Forest Decision Tree - SVM - Gradient Boosting - Voting Classifier. Calculated algorithm accuracy.
- User signup and login: Using this module will result in registration and login.
- User input: Using this module will result in predicted input.
- Prediction: final predicted shown

4. IMPLEMENTATION

ALGORITHMS:

KNN: The acronym KNN stands for "K-Nearest Neighbour". It is a machine learning algorithm that is supervised. The method can handle classification and regression problem statements. The sign 'K' represents the number of closest neighbours to a new unknown variable that must be predicted or categorised.

Naive Bayes: A probabilistic classifier, the Naive Bayes classification technique. It is based on probability models with high independence assumptions. The independence assumptions often have little effect on reality. As a result, they are seen as naïve.

A bagging classifier is an ensemble meta-estimator that fits base classifiers on random subsets of the original dataset and then aggregates their individual predictions (either by voting or average) to generate a final prediction. A meta-estimator of this kind is often used to

minimise the variance of a black-box estimator (for example, a decision tree) by incorporating randomization into its building mechanism and then constructing an ensemble from it.

Random Forest is a well-known machine learning algorithm that belongs to the supervised learning approach. It may be used to both classification and regression issues in machine learning. It is built on the notion of ensemble learning, which is a method that involves integrating several classifiers to solve a complicated issue and enhance the model's performance. "Random Forest is a classifier that comprises a number of decision trees on different subsets of the provided dataset and takes the average to enhance the predicted accuracy of that dataset," as the name implies. Instead than depending on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority vote of predictions.

Decision Tree: Decision trees use numerous methods to determine whether or not to divide a node into two or more sub-nodes. The development of sub-nodes promotes the homogeneity of the sub-nodes that arise. In other words, the purity of the node rises in relation to the target variable.

SVM: Support Vector Machine (SVM) is a common Supervised Learning method that is used for Classification and Regression tasks. However, it is mostly utilised in Machine Learning for Classification difficulties. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorising n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary.

Gradient Boosting: Gradient boosting is a machine learning approach that is often employed in regression and classification applications. It returns a prediction model in the form of an ensemble of weak prediction models, usually decision trees. When a decision tree is used as the weak learner, the resultant method is known as gradient-boosted trees, and it often beats random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as previous boosting techniques, but it extends the other approaches by enabling optimization of any differentiable loss function.

A voting classifier is a machine learning estimator that trains numerous base models or estimators and predicts based on the results of each base estimator. Aggregating criteria may be coupled voting decisions for each estimator output.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

5. EXPERIMENTAL RESULTS

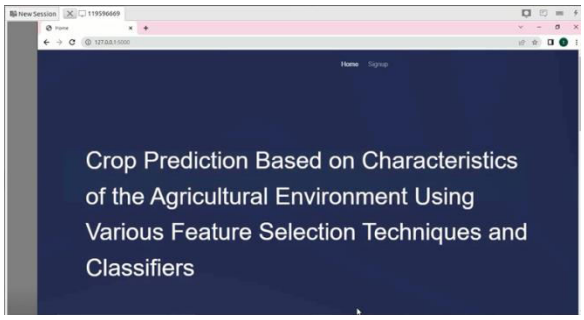


Fig.3: Home screen

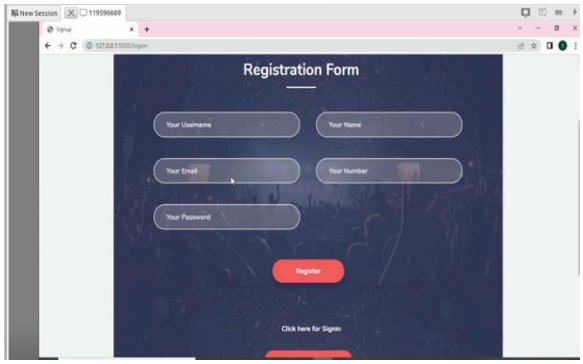


Fig.4: User registration

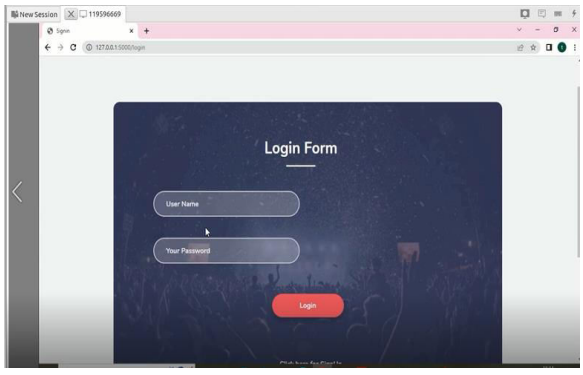


Fig.5: user login

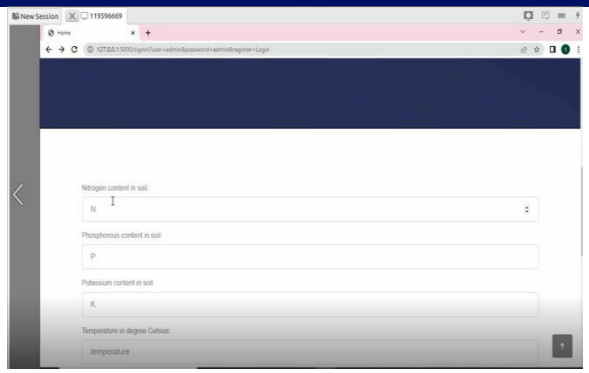


Fig.6: Main screen

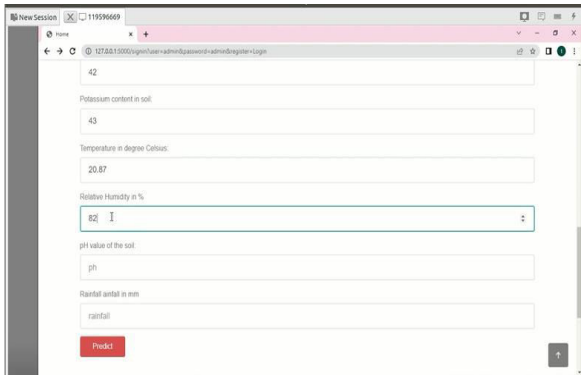


Fig.7: User input

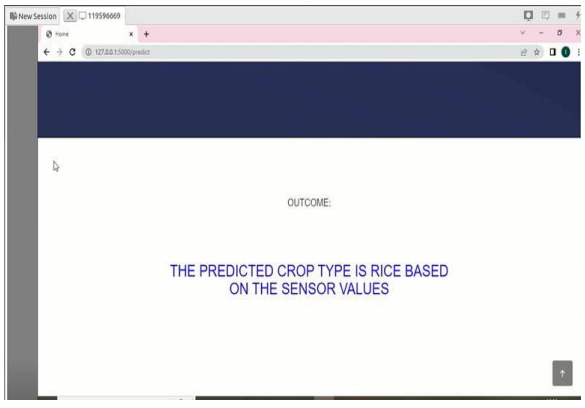


Fig.8: Prediction result

6. CONCLUSION

It is difficult to predict crops for cultivation in agriculture. This article employed a variety of feature selection and classification algorithms to estimate plant cultivation yield size. The findings show that an ensemble method outperforms the current classification technique in terms of prediction accuracy. Forecasting the area of grains, potatoes, and other energy crops may help farmers and countries plan the structure of their sowing. Modern forecasting approaches may result in substantial financial gains.

REFERENCES

[1] R. Jahan, ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[2] B. B. Sawicka and B. Krochmal-Marczak, ‘‘Biotic components influencing the yield and quality of potato tubers,’’ Herbalism, vol. 1, no. 3, pp. 125–136, 2017.

[3] B. Sawicka, A. H. Noaema, and A. Gáowacka, ‘‘The predicting the size of the potato acreage as a raw material for bioethanol production,’’ in Alternative Energy Sources, B. Zdunek, M. Olszáwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158–172.

[4] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ‘‘Biotic and abiotic factors influencing on the environment and growth of plants,’’ (in Polish), in Proc. Bioróżnorodność Środowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne, Puławy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192

[5] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ‘‘Response surface methodology: A retrospective and literature survey,’’ J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[6] D. K. Muriithi, ‘‘Application of response surface methodology for optimization of potato tuber yield,’’ Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[7] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ‘‘Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,’’ Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2014.

[8] J. R. Olędzki, ‘‘The report on the state of remotesensing in Poland in 2011–2014,’’ (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[9] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ‘‘Predicting of blue lupine yields based on the selected climate change scenarios,’’ Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.

[10] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.