Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28272

A Review of the Application of Supervised Learning in Agricultural Remote Sensing Monitoring: Yield Prediction

Yining Yang

School of Foreign Studies, Northeastern University at Qinhuangdao, Qinhuangdao, China
17721213023@163.com

Abstract. As the foundation of the national economy, agriculture is increasingly advancing toward informatization and intelligence, with remote sensing technology playing a vital role in large-scale crop monitoring and yield estimation. This paper systematically reviews the application of supervised learning in agricultural remote sensing, aiming to clarify its model characteristics, feature design methods, application logic, and key challenges—such as overreliance on single indicators like NDVI, spatial resolution limitations, and regional biases. The investigation, employing models such as Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, and Logistic Regression, demonstrates that although these techniques provide interpretability and reasonable accuracy, challenges persist in addressing sample scarcity, data imbalance, and the integration of multi-source data. The conclusion emphasizes the potential of integrating deep learning models (e.g., CNN, RNN) with traditional supervised learning to improve feature extraction, prediction accuracy, and generalization ability, providing a robust framework for future intelligent agricultural monitoring systems.

Keywords: Agricultural remote sensing, Supervised learning, Yield prediction, Machine learning

1. Introduction

As the bedrock of the national economy, agriculture's informatization and intelligent evolution have garnered widespread attention. Remote sensing, employing satellite or drone imagery, enables rapid, large-scale crop monitoring, playing a crucial role in contemporary agricultural management. Recent studies demonstrate that multi-temporal Sentinel-2 remote sensing imagery combined with NDVI grading and disaster information enables large-scale, long-term yield estimation for staple crops like corn, soybeans, and rice [1]. However, extracting practical information from remote sensing images has always been the core research issue. This review paper synthesizes supervised learning's theoretical applications in agricultural remote sensing, highlighting model characteristics, feature engineering, application logic, and existing research challenges. Current limitations include reliance on single indicators (e.g., NDVI for drought), insufficient satellite data resolution, and regional biases in crop disaster models [2]. Recent systematic reviews confirm that while NDVI is widely adopted (6 studies), temperature (24 studies) and rainfall (17 studies) remain the most frequently used features in crop yield prediction models [3]. This review paper synthesizes supervised learning

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ28272

applications in agricultural remote sensing, detailing model characteristics, feature engineering, application logic, and research challenges to inform future intelligent monitoring systems. Crop yield prediction is vital for food security amid population growth and climate change. Remote sensing and GIS, coupled with supervised machine learning, enable non-invasive yield estimation. This paper offers a methodological framework and future research directions for applying supervised learning to agricultural remote sensing yield prediction.

2. Supervised learning framework for yield prediction

2.1. Input features

In agricultural remote sensing for yield prediction, input feature selection is critical for model performance and accuracy. Commonly used input features include:

Normalized Difference Vegetation Index (NDVI): NDVI is a vegetation index calculated from satellite imagery that reflects vegetation's growth status and health. It is a critical nutritional status indicator for crops, calculated as NDVI=NIR−RNIR+RNDVI=NIR+RNIR−R (where NIR = near-infrared reflectance, R = red reflectance) [1]. Research has shown a significant correlation between NDVI and crop yield; a higher NDVI value typically indicates more vigorous vegetation growth and greater yield potential.

Temperature significantly impacts crop growth, with varying requirements across stages. Extremes harm yield. Data from stations/satellites informs models on environmental conditions. Precipitation is crucial for water supply; excesses/deficits affect yield. Data from stations/satellites improves yield variation explanations.

Soil Properties:Soil texture, fertility, and pH significantly affect crop yield. Hyperspectral and microwave remote sensing (e.g., MODIS, Sentinel-1) coupled with inversion models are used to monitor soil organic matter, moisture, and salinity. Deep-soil penetration and model generalizability remain challenges [2]. High soil fertility boosts crop yields, while soil texture impacts root growth and water retention. Soil properties from sampling or maps serve as input features, enhancing yield prediction models.

2.2. Output variables

In crop yield prediction, the choice of output variable depends on the specific research objectives and application scenarios. Standard output variables include:

Yield per Area: This metric directly measures crop yield per land unit, offering producers insights into land productivity for better planting and resource strategies.

Yield Grade: Categorizing yield into grades (e.g., high, medium, low) enables quick identification of key areas, informing agricultural policy and resource optimization.

2.3. Supervised learning process

The application process of supervised learning in crop yield prediction can be divided into the following steps:

Data preparation involves gathering remote sensing, meteorological, soil, and yield data. Time-series data, such as MODIS, are crucial due to the impact of interannual environmental variations on crop growth. Integrating multi-growth-stage and long-term remote sensing and meteorological data enhances growth monitoring precision [4]. Preprocessing (radiometric/geometric correction, cloud removal) ensures data usability.

DOI: 10.54254/2755-2721/2026.TJ28272

Feature Engineering: Extracting useful features from the collected raw data is a key step in improving model performance. Beyond the commonly used input features mentioned above, derivative features such as the rate of change of vegetation indices and cumulative values of temperature and precipitation can be calculated to enrich the feature set further.

Model Training: Select an appropriate supervised learning model, such as Decision Trees, Model training involves employing algorithms such as Random Forests, SVM, KNN, or Logistic Regression, with parameter tuning for optimal results. Performance is assessed via cross-validation using MSE, R^2, or classification accuracy. Optimization includes adjusting model structure, feature selection, or ensemble methods to enhance prediction accuracy and generalization.

3. Key supervised learning models

3.1. Decision trees and Random Forests

Decision tree-based crop mapping, while integrating multi-dimensional features (NDVI, soil, temperature), shows higher prediction errors (e.g., 12.4% for corn) than temporal models, limited by remote sensing data and sparse ground truth [2]. They improve model accuracy and stability by constructing multiple decision trees and aggregating their predictions. Empirical studies show that while Random Forests effectively integrate multi-dimensional features (NDVI, soil type, temperature), their prediction errors for corn (12.4%), soybeans (13.7%), and rice (16.1%) remain higher than temporal models like LSTM [1]. Random Forests excel at handling high-dimensional and noisy data, effectively avoiding overfitting. Random Forests in yield prediction leverage multiple inputs and trees via random sampling, improving adaptability to varied conditions. Studies show Random Forests achieve >85% accuracy in agricultural remote sensing yield prediction.

3.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) are a supervised learning model based on statistical learning theory. Their goal is to find an optimal hyperplane that separates data points of different classes. In agricultural remote sensing yield prediction, SVM can separate feature vectors of high-yield and low-yield areas to achieve yield grade classification or yield per unit area prediction. SVM excels in classifying small-sample, high-dimensional data by using kernel functions to map data into higher-dimensional spaces for non-linear problem-solving. In yield prediction, SVM effectively identifies non-linear relationships between multi-dimensional features like NDVI, soil fertility, and meteorological data for accurate predictions. However, SVM's training is time-intensive and sensitive to parameter selection, necessitating optimization via cross-validation.

3.3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a similarity-based, non-parametric supervised learning method. Its core idea is to predict the label or value of a target point based on the labels of the K closest data points (neighbors) in the training dataset. In agricultural remote sensing for yield prediction, KNN estimates a target plot's yield by calculating similarity metrics (e.g., Euclidean distance) to K plots with known yields and averaging their yield data [6]. KNN's simplicity and ease of implementation without data modelling assumptions are its advantages. However, K value and distance metric greatly affect its performance. Its computational performance is low for huge datasets [5]. Additionally, KNN is sensitive to noise and outliers in the data, requiring data preprocessing and feature normalization.

DOI: 10.54254/2755-2721/2026.TJ28272

3.4. Logistic Regression

Logistic Regression is a supervised learning model for binary or multi-class classification problems. It maps input features to probability values within the [0,1] interval using a logistic function to achieve classification prediction. In agricultural remote sensing yield prediction, Logistic Regression can classify yield into high, medium, and low grades, predicting the yield grade of a target plot by calculating the probability value for each grade, Model parameters can be optimised using maximum likelihood estimation. Its merits are model simplicity, interpretation ease, and quick training and prediction. Logistic regression cannot predict complex nonlinear interactions because it presupposes linearly independent input features [7].. Therefore, feature engineering or combination with other models may be necessary when handling high-dimensional non-linear data

3.5. Model comparison

Supervised learning algorithms for agricultural remote sensing yield prediction offer pros and cons depending on the application. Decision Trees and Random Forests are appropriate for non-linear relationships and high-dimensional data due to their interpretability and noise resilience, however Random Forests take longer to train. SVM performs well on small-sample, high-dimensional data but is sensitive to parameter selection and has a long training time. KNN is simple and easy to implement, but computationally inefficient and sensitive to noise and outliers. Logistic Regression is interpretable and straightforward but has weak modeling capability for non-linear relationships.

Decision Trees and Logistic Regression are simpler than Random Forests and SVM. KNN and Logistic Regression train quicker than Random Forests and SVM. KNN and SVM are weaker than Decision Trees and Logistic Regression in interpretability.

Overall, selecting the appropriate supervised learning model requires a comprehensive consideration of data characteristics, task requirements, and model performance metrics. For agricultural remote sensing yield prediction tasks, choosing a suitable model based on the specific application scenario and data features or employing model fusion methods to enhance prediction accuracy and model generalization capability is recommended. In global comparative analyses, Neural Networks (including deep learning variants) are the most frequently applied model (54% of studies), followed by Linear Regression (28%) and Random Forests (24%) [3].

4. Research challenges

4.1. Sample label construction

Constructing sample labels is a critical step in supervised learning for crop yield prediction, but it faces numerous practical challenges. Firstly, in multi-plot monitoring, the vast number and wide distribution of plots make precise yield labeling for each plot labor-intensive and costly. Large farms or cross-regional agricultural monitoring initiatives make it impossible to measure and record production for every plot, resulting in missing yield labels. Yield data may vary by region and year, making label uniformity and accuracy difficult. Poor label data can affect model training and prediction accuracy. Labelled yield data at scale is scarce; 32% of research see "insufficient data availability" as a major restriction, especially for sophisticated model training [4].

DOI: 10.54254/2755-2721/2026.TJ28272

4.2. Data imbalance

Supervised agricultural yield prediction often suffers from data imbalance. Actual datasets have few high-yield samples but many low- or medium-yield samples. This imbalanced data distribution biases the model towards the majority class during training, making minority class samples (e.g., high-yield samples) unpredictible. Due to soil and climate conditions, high-yield plots may be rare in some agricultural regions. Due to many low-yield samples, the model may struggle to detect high-yield plots. To balance datasets, researchers use oversampling, undersampling, or synthetic data synthesis. However, these strategies may add biases or lose information.

4.3. Integration of multi-source heterogeneous data

Crop yield prediction often requires integrating heterogeneous data from diverse sources, such as remote sensing imagery, soil data, and meteorological data. These data types exhibit significant differences in format, resolution, and temporal scale, posing substantial challenges for integration. For instance, remote sensing imagery might have high spatial resolution but low temporal resolution, while meteorological data might have high temporal resolution but low spatial resolution. Effectively fusing these data with different resolutions to leverage the strengths of each type is a current research challenge. Furthermore, noise and errors from different data sources may interact, increasing the complexity of data integration. Researchers must develop more advanced data fusion techniques to improve data usability and model prediction performance.

4.4. Model transferability and generalization

The transferability and generalization capability of supervised learning models in crop yield prediction is another significant issue in current research. Many models perform well on their training datasets but experience significant performance degradation when applied across different regions or years. This is exacerbated by regional biases in existing models—e.g., disaster monitoring focuses predominantly on winter wheat and rice, with limited applicability to orchards or southwestern China [3]. This transferability problem limits the broader application of these models. Researchers must consider more environmental factors and data diversity to enhance model generalisation, adopt more complex model architectures, or use ensemble learning methods. Additionally, introducing transfer learning techniques can partially alleviate model transferability issues, although further research and validation are still required.

5. Integration with deep learning models

As agricultural remote sensing data volume and processing capacity rise, supervised learning with deep learning models will be crucial for agricultural production prediction. Deep learning models like CNNs and RNNs excel at processing large-scale image and time-series data. CNNs can automatically extract complicated features from remote sensing pictures, whereas RNNs can handle meteorological time-series data. Combining deep learning models with supervised learning models lets deep learning extract features and supervised learning interpret them. Recent empirical research support this hybrid strategy. In Ukraine, Kussul suggested a multilayer DL architecture for crop categorisation utilising multitemporal Landsat-8 and Sentinel-1A images. CNN ensemble improved accuracy to 94.6%, surpassing Random Forests (88.7%) and Multilayer Perceptrons (92.7%). CNNs considerably improved crop type discrimination, with maize accuracies of 94.6% and soybeans of 86.9%. CNN can represent hierarchical spectral-temporal properties, overcoming RF and MLP's

DOI: 10.54254/2755-2721/2026.TJ28272

limitations in complicated agricultural landscapes [8]. UNet3+ (with CBAM attention mechanism enhancing segmentation Dice coefficient from 0.9105 to 0.9403) and LSTM models have shown >90% prediction accuracy for important field crops, confirming this hybrid technique [1]. Furthermore, transfer learning techniques offer new approaches to address model transferability and generalization challenges. By pre-training deep learning models on large-scale datasets and then transferring them to specific agricultural regions or tasks, the model's adaptability to new environments can be significantly enhanced. For example, a deep learning model trained on one region can be quickly adapted to the yield prediction task in another region through fine-tuning, reducing reliance on large amounts of labeled data.

Better multi-source heterogeneous data integration into deep learning frameworks can be explored in future research. Using novel data fusion methods to merge remote sensing pictures, soil data, and meteorological data to provide more detailed feature representations. This enhances model prediction and agricultural decision-making data.

6. Conclusion

This paper systematically reviews the theoretical application of supervised learning in agricultural remote sensing for crop yield prediction. By analyzing the models, features, application mechanisms, and performance metrics, we have provided a comprehensive overview of the current state of the art. Our findings indicate that basic models such as decision trees, random forests, and KNN demonstrate strong recognition capabilities and interpretability under remote sensing imagery, making them suitable for agricultural applications. Future research should explore integrating deep learning models to enhance prediction accuracy and model generalization. Combining the feature extraction capabilities of deep learning with the interpretability of supervised learning could lead to more robust and efficient yield prediction systems. Moreover, developing advanced data fusion techniques to integrate remote sensing, soil, and meteorological data could provide a more comprehensive feature set for model training.

In conclusion, supervised learning offers a viable theoretical pathway for agricultural remote sensing and has the potential to contribute significantly to the intelligent upgrading of global agriculture. Further research and development in this area are essential to address existing challenges and leverage these technologies' full potential for food security and agricultural management.

References

[1] Cong, H. (2024). Research on crop yield estimation models based on remote sensing images (Master’s thesis, Jilin Agricultural University). https: //doi.org/10.27163/d.cnki.gjlnu.2024.000572

[2] Zhuang, J. Y., Bao, W. J., & Su, W. Z. (2024). Current status and prospects of agricultural remote sensing applications. Agricultural Outlook, 20(04), 68-74.

[3] Van K., T., Ayalew, K., & Cagatay, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 175, 105709. https: //doi.org/10.1016/j.compag.2020.105709

[4] Wang, P. X., Tian, H. R., Zhang, Y., Han, D., Wang, J., & Yin, M. (2022). Research progress on crop growth monitoring and yield estimation based on deep learning. Transactions of the Chinese Society of Agricultural Machinery, 53(02), 1-14.

[5] Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. Scientific World Journal, 2014, 509429. https: //doi.org/10.1155/2014/509429

[6] Shah, A., Dubey, A., Hemnani, V., Gala, D., & Kalbande, D. R. (2018). Smart farming system: Crop yield prediction using regression techniques. In H. Vasudevan, A. Deshmukh, & K. Ray (Eds.), Proceedings of

DOI: 10.54254/2755-2721/2026.TJ28272

International Conference on Wireless Communication (Lecture Notes on Data Engineering and Communications Technologies, vol. 19). Springer, Singapore. https: //doi.org/10.1007/978-981-10-8339-6_6

[7] Ma, Y., Zhang, Z., Kang, Y., & Özdoğan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sensing of Environment, 259, 112408. https: //doi.org/10.1016/j.rse.2021.112408

[8] Nataliia, K., Mykola, L., Sergii, S., & Andrii, S. (2017). Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(10), 1896-1900. https: //doi.org/10.1109/lgrs.2017.2681128