



Machine Learning-Based Prediction of Tomato Yield in Greenhouse Environments

M'hamed Mancer, Labib Sadek Terrissa, and Soheyb Ayad*

LINFI Laboratory, University of Biskra, Algeria.

* Corresponding author: s.ayad@univ-biskra.dz

Abstract. The agricultural sector heavily relies on accurate crop yield predictions, providing farmers with crucial information to manage their crops, allocate resources efficiently, and plan market strategies. This article proposes a novel approach utilizing a Stacked Ensemble Model for predicting tomato crop yield in greenhouse environments. A comprehensive dataset encompassing various factors related to greenhouse climate, crop parameters, and production was used for training and evaluating the models. Comparative analysis with other advanced regression models, including K-Nearest Neighbors (KNN), Random Forest, and Light-GBM, demonstrated the superior performance of the Stacked Ensemble Model, highlighted by the highest R2 value (0.896) and the lowest mean squared error (MSE) of 0.008. These results signify heightened accuracy and a close alignment between predicted and actual values. Our proposed system empowers farmers with the ability to accurately predict tomato yield, enabling them to mitigate risks, optimize harvest schedules, and effectively meet market demands.

Keywords: Tomato yield prediction · Machine Learning · Smart agriculture · Greenhouse environments · Precision agriculture.

1 Introduction

The tomato, a versatile and nutritious fruit, is a cornerstone of global agriculture. Its widespread consumption and significant economic value make it a crucial crop for food security and economic stability worldwide [1]. However, accurately anticipating tomato yields remains a challenge for farmers, hindering their ability to strategize growth, optimize resource allocation, and maximize profitability [2].

Traditional methods of yield prediction often rely on manual observation and expert knowledge, which can be subjective and prone to inaccuracies, particularly in complex growing environments like greenhouses [3].

Enter smart agriculture. This revolutionary approach integrates cutting-edge technologies like Artificial Intelligence (AI), the Internet of Things (IoT), Blockchain (Distributed Ledger), and robotics to transform agricultural practices [4] [5]. Machine Learning (ML) plays a central role in smart agriculture by empowering farmers with data-driven insights. These insights enable them to make informed

© The Author(s) 2024

C. A. Kerrache et al. (eds.), *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, Advances in Intelligent Systems Research 184,

https://doi.org/10.2991/978-94-6463-496-9_10

decisions about crop management, predict yields with greater accuracy, and optimize resource utilization, ultimately promoting sustainable agricultural practices that benefit both the environment and their bottom line [6] [7].

This research aims to contribute to this advancement by proposing a precise model for estimating tomato yields. Unlike traditional methods, this model considers a multitude of factors that influence yield, including the complex interplay of greenhouse climate conditions, stem characteristics, and historical harvest information. To achieve this goal, the study leverages the Stacked Ensemble Model as the core analytical approach.

Through this research, we aim to provide farmers with a robust tool for accurately predicting tomato yield, enabling them to mitigate risks, optimize harvest schedules, and effectively meet market demands.

2 Literature Review

In this section, we delve into the realm of tomato yield prediction, a domain that has garnered relatively fewer studies compared to other crops. Nevertheless, we aim to explore and present a range of pertinent research endeavors that focus on yield prediction across different crops, employing diverse methodologies and approaches.

For instance, the authors of this study [8] presented research on winter wheat yield prediction using machine learning (ML) and deep learning approaches. They utilized a large dataset comprising weather, soil, and crop phenology variables from 271 counties across Germany, spanning the period from 1999 to 2019. The objective was to analyze the efficacy of various models, including deep neural networks (DNNs), convolutional neural networks (CNNs), decision trees, random forests, XGBoost, and linear regression. Achieving 7 to 14% lower RMSE, 3 to 15% lower MAE, and improved correlation coefficients by 4 to 50%, the proposed CNN model outperformed all other models in predicting winter wheat production.

Another study [9] developed a novel approach for the early estimation of tomato yield using Decision Tree Ensembles (DTE) and information obtained from Unmanned Aerial Vehicles (UAVs). The results showed that the DTE-Bag model could predict tomato yield with an accuracy of 92.5%. The authors believe that their proposed method may be utilized to enhance the predicting of tomato yield and assist farmers in making informed decisions.

Furthermore, the authors of this paper [10] developed a transformer-based model for accurately predicting rice yield using satellite and climatic indicators. The model outperformed four other ML models (LASSO, RF, XGBoost, and AtLSTM), achieving the highest R^2 of 0.78, lowest RMSE of 0.44 t/ha, lowest MAPE of 16.56%, and greatest accuracy of 0.72. Although the proposed model effectively predicts rice yield by considering various factors, the study has some limitations, such as the lack of consideration of other features like soil properties, tillage, and fertilization rates, which are also closely related to rice yield.

In another study [11], the outputs of a biophysical model and an ML model are combined to propose a unique method for predicting tomato yields in a greenhouse. The Tomgro model is a commonly used biophysical model for predicting tomato yields based on environmental parameters such as temperature, humidity, and light. Convolutional neural network-recurrent neural network (CNN-RNN) modeling is used in this method to predict future yields, trained using historical yield data and environmental parameters. Among all the techniques, the fusion strategy yielded the most accurate prediction results, with RMSEs, R^2 s, and Nash-Sutcliffe efficiency (NSE) mean and standard deviations of 17.69 ± 3.47 .

Additionally, using available climatic data such as temperature, precipitation, and solar radiation, this study [12] provided a simple and easy approach for predicting national wheat yields. The study found that the Random Forest model was the best-performing model, with an RMSE of 9.1

Moreover, the purpose of this study [13] was to compare different rice yield prediction models, such as multiple linear regression (MLR), RF, and the traditional model (TR), based on the most important agronomic traits (plant number per m^2 and plant height). The research was carried out in China's Jilin Province. The findings revealed that the RF model outperformed the other models in terms of accuracy and robustness.

This research [2] evaluates the effectiveness of various regression models—Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Lasso Regression, and Linear Regression—in predicting tomato yields across different datasets. Dataset 3 is identified as the most comprehensive and accurate. While RF requires more computational time, it demonstrates the highest predictive accuracy. KNN and Lasso Regression perform well with lower computational costs.

Through an extensive review of existing literature, our study endeavors to advance tomato yield prediction by enhancing predictive models. We aim to elucidate the outcomes of the proposed system in terms of yield, explicate the results of the Stacked Model, and identify potential enhancements.

3 PROPOSED METHOD

We have developed a Stacked Ensemble Model aimed at predicting the daily tomato yield within a greenhouse environment. The overarching framework of our proposed approach, as depicted in Figure 1, comprises both online and offline stages. The offline phase involves constructing the prediction model, encompassing data preprocessing and Stacked training. Subsequently, the online phase entails employing the trained Stacked Model to predict new instances and assess their performance.

The Stacked Ensemble Model is specifically chosen due to its robustness in handling multicollinearity within datasets characterized by correlated features. Ensemble methods effectively mitigate overfitting concerns by combining predictions from multiple models, enhancing the generalization to new and unseen

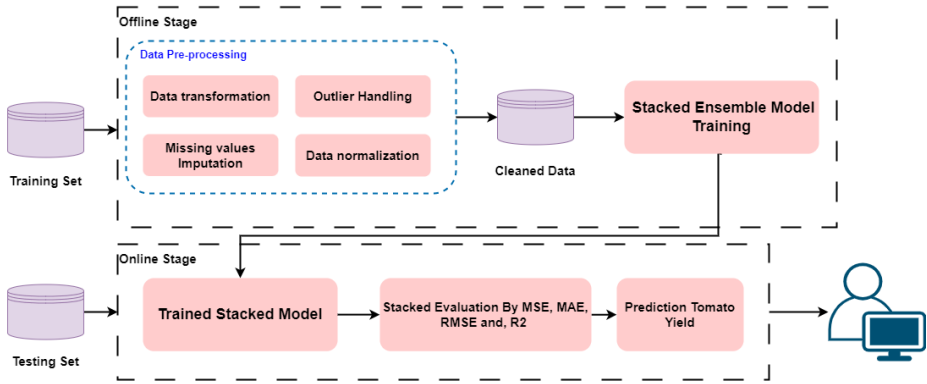


Fig. 1. The general architecture of the proposed system.

data [14]. This choice aligns with our goal of achieving accurate and reliable predictions of daily tomato yield in a greenhouse environment.

The efficacy of the Stacked Model in managing multicollinearity and preventing overfitting will be demonstrated and substantiated in the subsequent *Results and Discussion* section.

3.1 Dataset description

The dataset utilized in this research was obtained from the Autonomous Greenhouse Competition (AGC) (second edition) [15]. It provides a comprehensive collection of data that captures various aspects of tomato cultivation and greenhouse management recorded during the competition.

The dataset consists of six collections, each containing six sub-datasets. Our study utilized three specific datasets: the *Greenhouse climate*, *Crop parameters*, and *Production* datasets. These selections were made because they directly relate to the growth and yield prediction of tomatoes. Table 1 provides a detailed explanation of the datasets used.

The *Greenhouse climate* dataset includes detailed information on the greenhouse climate during tomato plant growth. It contains 47,808 rows of data gathered over 166 days, with each row representing climatic readings collected at 5-minute intervals. The *Crop parameters* dataset includes detailed information on tomato plant growth, such as stem growth, stem thickness, stem density, and plant density. It contains 23 rows of data, with each row representing measurements collected at weekly intervals. Finally, the *Production* dataset includes detailed information on tomato plant harvests, such as total production, the number of harvested fruits, and the weight of harvested fruits. It contains 24 rows of data, with each row representing measurements collected on the harvest date. In this research, we were especially interested in the 'total production' and did not use data related to the number or weight of the fruits.

Table 1. Dataset Description

Dataset	Time	Feature	Description	Unit
Greenhouse Climate	5 min	Tair	Greenhouse air temperature	°C
		Rhair	Greenhouse relative humidity	%
		CO2air	CO2 in greenhouse	ppm
		co2_dos	CO2 dosing	kg/ha hour
		Tot_PAR	Total inside PAR (Sun + HPS + LED)	$\mu\text{mol}/\text{m}^2 \text{ s}$
		EC_drain	Drain EC	dS/m
		pH_drain	Drain pH	-
		Cum_irr	Cumulative irrigation per day	$\text{L}/\text{m}^2 \text{ day}$
Crop Parameters	Weekly	Stem_thick	Stem thickness	mm
		Stem_elong	Stem growth per week	cm/week
		Stem_dens	Stem density	Stems/ m^2
		Plant_dens	Plant density	Plants/ m^2
Production	at harvest date	prod	Total tomato production	kg/m^2

3.2 Data Preprocessing

Data transformation From Table 1, we observe that the data collection intervals for different features are not uniform. To ensure compatibility and usability of all features within a single model, we standardized the collection intervals to a daily timeframe.

For the *Greenhouse climate* features collected at 5-minute intervals, we transformed the data by calculating the mean value for each day. Similarly, the data for the *Crop parameters* characteristics, which were gathered weekly, were re-sampled to a daily frequency. The same resampling technique was applied to the *Production* data to obtain daily values (Fig. 2).

Handling Missing Values Handling missing values is a critical phase in data preprocessing due to its significant impact on subsequent analyses. Various factors such as sensor malfunction, data recording discrepancies, or incomplete data collection can lead to missing values [16]. In this study, we addressed missing values using a straightforward yet effective method known as imputation. Imputation involves substituting missing data with values derived from statistical measures. Specifically, we used median values to impute missing data points. This approach ensures data integrity and facilitates comprehensive analysis by reducing the impact of missing values on the dataset.

Handling Outliers Dealing with outliers is another crucial aspect of data processing that can significantly affect subsequent analyses. Outliers may arise due to measurement errors, experimental anomalies, or inherent dataset fluctuations. In this study, we employed robust methodologies using statistical measures such as the interquartile range (IQR) to handle outliers. Specifically, we identified and

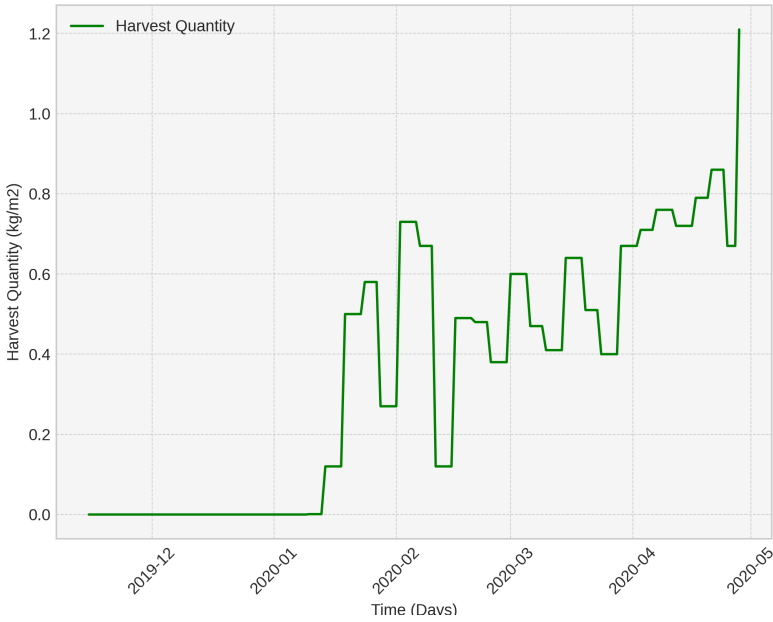


Fig. 2. Daily Harvest of Tomato Crop.

treated outliers by replacing them with the median value. This approach aims to mitigate the influence of outliers on analyses while preserving dataset integrity.

Data Normalization Data normalization aims to standardize the range and distribution of features within the dataset, particularly when dealing with features exhibiting diverse measurement units or ranges [17]. In this study, we utilized the MinMax scaler for data normalization. This method effectively eliminates biases arising from differing feature scales, facilitating fair comparison and analysis of the dataset.

3.3 Machine Learning Models

Within this section, we delineate the machine learning (ML) models employed in our study. Our primary focus lies in the proposition and implementation of a Stacked Ensemble Model. Additionally, we leverage K-Nearest Neighbors Regression (KNN), Gradient Boosting Machines using LightGBM, and Random Forest to facilitate a comprehensive comparison and evaluation of our results. This selection enables us to assess the predictive performance and efficacy of the proposed Stacked Ensemble Model against both established and advanced models, thereby enhancing the robustness of our analysis.

Stacked Ensemble Model: The Stacked Ensemble Model employs a model stacking approach that combines multiple base learners (Ridge, Random Forest, and XGBoost) using a meta-learner. This method aims to enhance predictive performance by leveraging the strengths of each base learner and mitigating their weaknesses [14].

K-Nearest Neighbors Regression (KNN): KNN is a simple yet powerful algorithm used for regression tasks. It predicts the target variable by identifying the k closest neighbors to a given data point and computing the average or weighted average of those neighbors' target values. [20].

LightGBM: LightGBM is a highly efficient gradient-boosting framework that employs tree-based learning algorithms. It is designed for distributed computing and can handle large-scale data efficiently, making it suitable for high-dimensional datasets [18].

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. This approach helps to improve accuracy and control overfitting [19].

Through comprehensive evaluation and comparison with KNN, LightGBM, and Random Forest, we aim to showcase each model's performance metrics, including mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R-squared). These metrics will provide empirical evidence supporting the effectiveness of these models in accurately predicting tomato yield within the greenhouse environment. The results will highlight the predictive power and reliability of the Stacked Ensemble Model, demonstrating its superiority over traditional and advanced individual models.

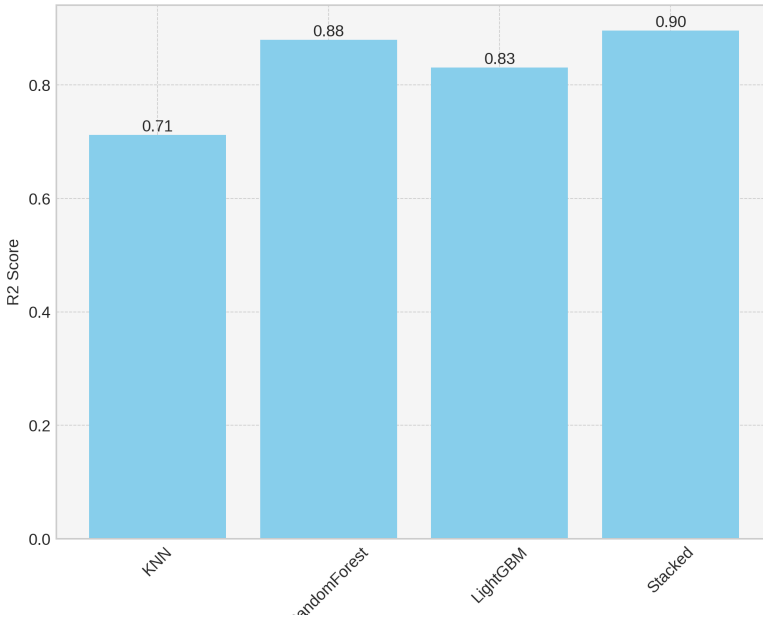
4 Results and Discussion

In this section, we present and discuss the results obtained from our study on tomato yield prediction using various machine-learning models. The models include K-Nearest Neighbors (KNN), Random Forest, LightGBM, and a Stacked Ensemble Model. To ensure robust evaluation, the dataset was partitioned into a training set (80%) and a testing set (20%). Performance metrics after hyperparameter tuning are summarized in Table 2, with visual representations of the results shown in Figs. 3, 4, and 5.

According to the R² comparison (Fig. 3), the Stacked Ensemble Model achieved the highest R² value of 0.896, surpassing RandomForest at 0.884, LightGBM at

Table 2. Performance Evaluation of Tomato Yield Prediction Models

Model	MSE	MAE	RMSE	R2
KNN	0.0226	0.11	0.15	0.712
RandomForest	0.009	0.046	0.095	0.884
LightGBM	0.013	0.083	0.114	0.831
Stacked	0.008	0.065	0.09	0.896

**Fig. 3.** Comparison of R2 Scores.

0.831, and KNN at 0.712. These results indicate the Stacked Ensemble Model's superior capacity to explain variance within the data, signifying a stronger fit than the other models.

The higher R2 value of the Stacked Ensemble Model suggests its effectiveness in capturing the underlying patterns of the data. This model's ability to combine predictions from multiple base learners, such as Ridge, Random Forest, and XGBoost, enhances its performance by leveraging the strengths of each model and mitigating its weaknesses. In terms of error metrics (Fig. 4), the Stacked Ensemble Model demonstrated the lowest mean squared error (MSE) at 0.008, indicating superior prediction accuracy with fewer errors. The RandomForest, LightGBM, and KNN models reported slightly higher MSE values

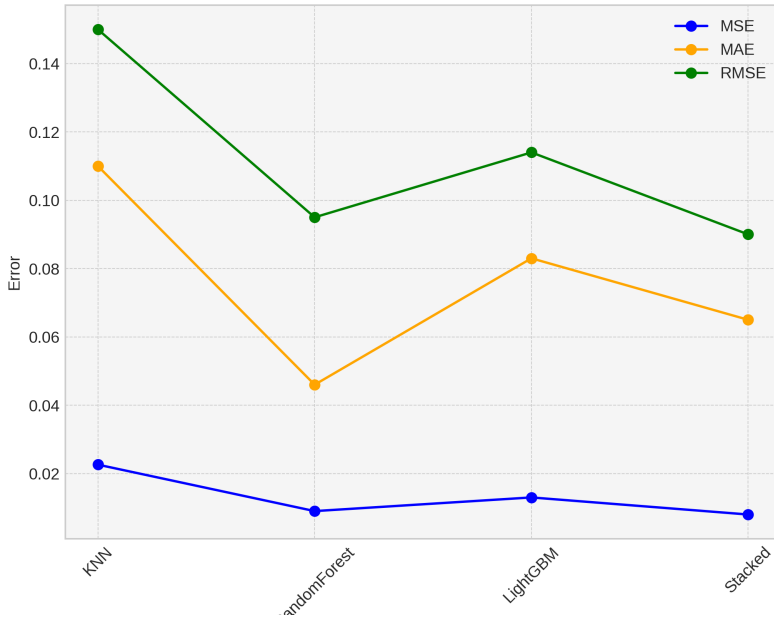


Fig. 4. Comparison of Error Metrics.

of 0.009, 0.013, and 0.0226, respectively. Additionally, the Stacked Ensemble Model outperformed others with the lowest root mean squared error (RMSE).

The Stacked Ensemble Model's superior performance in minimizing errors can be attributed to its ensemble approach. Combining multiple models effectively captures a broader range of data patterns and reduces the risk of overfitting, leading to more accurate predictions and better generalization to new data. Fig. 5 provides a comprehensive visual analysis of the alignment between the predicted and actual values, allowing for a full evaluation of prediction accuracy. Ideally, the predicted values should coincide with the diagonal line, indicating a high level of agreement between the model's predictions and the actual values.

The Stacked Ensemble Model's predictions align closely with the diagonal line, demonstrating its superior predictive capability and closer approximation to actual values compared to the KNN, LightGBM, and RandomForest models. The slight deviations observed in the KNN, LightGBM, and RandomForest models highlight their relatively lower accuracy in comparison.

The results of this study demonstrate the benefits of using a Stacked Ensemble Model for tomato yield prediction. The ensemble approach's ability to integrate multiple learning algorithms results in a model that outperforms individual models like KNN, LightGBM, and RandomForest in terms of R^2 , MSE, MAE, and RMSE.

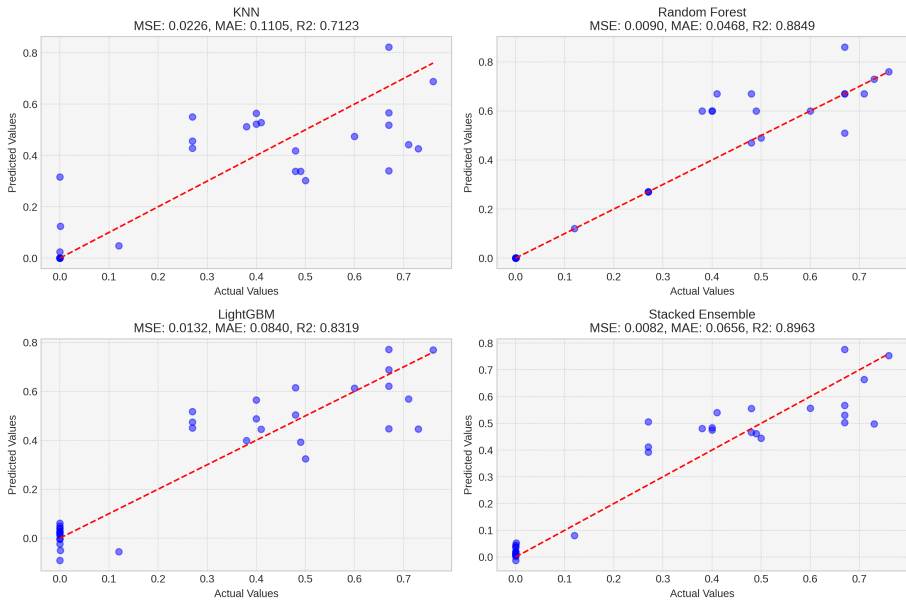


Fig. 5. Comparison of Predicted and Actual Values for Models.

The ensemble method's strength lies in its robustness and flexibility. By combining models that capture different aspects of the data, the ensemble model can effectively balance bias and variance, leading to improved generalization and prediction accuracy. This is particularly advantageous in complex agricultural datasets where variability and non-linear relationships are common.

Furthermore, the lower error metrics achieved by the Stacked Ensemble Model underscore its potential for practical applications in precision agriculture. Accurate yield predictions can significantly aid farmers in making informed decisions about resource allocation, crop management, and market strategies, ultimately enhancing productivity and sustainability.

5 Conclusion

This paper introduced a machine-learning strategy employing various models to predict tomato crop yield in greenhouse conditions. The models evaluated include K-Nearest Neighbors (KNN), Random Forest, LightGBM, and a Stacked Ensemble Model. Our results demonstrate that the Stacked Ensemble Model outperforms the other models, achieving the highest R2 value of 0.896 and the lowest mean squared error (MSE) of 0.008. These metrics indicate superior accuracy and a closer alignment between predicted and actual values.

The Stacked Ensemble Model's success can be attributed to its ability to integrate multiple base learners, capturing diverse patterns within the data and

reducing the risk of overfitting. This model's enhanced predictive performance underscores its potential as a robust tool for farmers, aiding in optimized harvest planning, meeting market demands, and advancing agricultural practices.

Future research could explore alternative machine learning methodologies, including deep learning approaches, to further enhance predictive performance. Additionally, validating this system across diverse geographical areas and varying agricultural conditions would enhance its applicability and credibility.

References

1. Tomato Land & Water <https://www.fao.org/land-water/databases-and-software/crop-information/tomato/en/>. Accessed 07-Jul-2023
2. Mancer, M., Terrissa, L., Ayad, S. & Laouz, H. Tomato Crop Forecasting: A Comparative Analysis of Regression Models. *2024 ASU International Conference In Emerging Technologies For Sustainability And Intelligent Systems (ICETSSIS)*. pp. 649-653 (2024)
3. Muruganatham, P., Wibowo, S., Grandhi, S., Samrat, N. & Islam, N. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*. **14**, 1990 (2022)
4. Mancer, M., Akram, K., Barka, E., Okba, K., Sihem, S., Harous, S., Athamena, B. & Houhamdi, Z. Blockchain Technology for Secure Shared Medical Data. *2022 International Arab Conference On Information Technology (ACIT)*. pp. 1-6 (2022)
5. Mancer, M., Terrissa, L., Ayad, S. & Laouz, H. A Blockchain-based Approach to securing data in smart agriculture. *2022 International Symposium On INnovative Informatics Of Biskra (ISNIB)*. pp. 1-5 (2022)
6. Raghuvanshi, A., Singh, U., Sajja, G., Pallathadka, H., Asenso, E., Kamal, M., Singh, A. & Phasinam, K. Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming. *Journal Of Food Quality*. **2022** pp. 1-8 (2022)
7. Mancer, M., Terrissa, L., Ayad, S., Laouz, H. & Zerhouni, N. Advancing Crop Recommendation Systems Through Ensemble Learning Techniques. *The Proceedings Of The International Conference On Smart City Applications*. pp. 45-54 (2023)
8. Srivastava, A., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T. & Rahimi, J. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*. **12**, 3215 (2022)
9. Lillo-Saavedra, M., Espinoza-Salgado, A., Garcia-Pedrero, A., Souto, C., Holzapfel, E., Gonzalo-Martin, C., Somos-Valenzuela, M. & Rivera, D. Early estimation of tomato yield by decision tree ensembles. *Agriculture*. **12**, 1655 (2022)
10. Liu, Y., Wang, S., Chen, J., Chen, B., Wang, X., Hao, D. & Sun, L. Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method. *Remote Sensing*. **14**, 5045 (2022)
11. Gong, L., Yu, M., Cutsuridis, V., Kollias, S. & Pearson, S. A Novel Model Fusion Approach for Greenhouse Crop Yield Prediction. *Horticulturae*. **9**, 5 (2022)
12. Júnior, R., Olivier, L., Wallach, D., Mullens, E., Fraisse, C. & Asseng, S. A simple procedure for a national wheat yield forecast. *European Journal Of Agronomy*. **148** pp. 126868 (2023)
13. Liu, B., Liu, Y., Huang, G., Jiang, X., Liang, Y., Yang, C. & Huang, L. Comparison of yield prediction models and estimation of the relative importance of main agronomic traits affecting rice yield formation in saline-sodic paddy fields. *European Journal Of Agronomy*. **148** pp. 126870 (2023)

14. Zhai, B. & Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average PM_{2.5} concentrations in Beijing, China. *Science Of The Total Environment*. **635** pp. 644-658 (2018)
15. Hemming, S., Zwart, H., Elings, A., Petropoulou, A. & Righini, I. Autonomous Greenhouse Challenge, Second Edition (2019). (4TU.ResearchData,2020), https://data.4tu.nl/articles_/12764777/2
16. Saar-Tsechansky, M. & Provost, F. Handling missing values when applying classification models. (Journal of Machine Learning Research,2007)
17. Quackenbush, J. Microarray data normalization and transformation. *Nature Genetics*. **32**, 496-501 (2002)
18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems*. **30** (2017)
19. Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R. & Friedman, J. Random forests. *The Elements Of Statistical Learning: Data Mining, Inference, And Prediction*. pp. 587-604 (2009)
20. Song, Y., Liang, J., Lu, J. & Zhao, X. An efficient instance selection algorithm for k nearest neighbour regression. *Neurocomputing*. **251** pp. 26-34 (2017)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

