Predicting Agricultural Success with Machine Learning

K. Chinnathambi[1], Rajyalakshmi Kandimalla[2] and Srihari Rayala[3]
{ drchinnathambik@veltech.edu.in[1], rajyalakshmi12@gmail.com[2], rayalasrihari9@gmail.com[3] }

Assistant Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India[1]
Student, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India[2, 3]

Abstract. This report applies the KNN algorithm, a simple yet effective machine learning method, to predict agricultural success. Through historical agricultural data, weather information, soil conditions, and other related data, the KNN model forecasts crop yields with high precision. This forecast will serve as a guide for farmers and industries in the agriculture sector to take data driven decisions to efficiently manage resources and increase productivity. For this purpose, we use KNN algorithm, which is a very straight forward and effective machine learning algorithm for classification and regression studies, to establish a model for crop yield prediction under different environmental and agricultural conditions. The research demonstrates the application of machine learning mainly K-nearest neighbours in improving agricultural practice and ensuring food security.

Keywords: Data-Driven Decisions, Food Security, Regression Tasks, Classification Tasks.

1 Introduction

Agriculture is the foundation of global food security, and accurate crop yield estimation is essential for improving efficiency and resource utilization. Machine learning (ML) methods, particularly the K-Nearest Neighbours (KNN) algorithm, have demonstrated strong potential in predicting yields by incorporating environmental factors such as weather conditions, soil characteristics, and historical records. Previous studies confirm that ML-based approaches can support farmers with reliable, data-driven decision-making.

KNN is a non-parametric, instance-based learning algorithm valued for its simplicity and effectiveness. It is frequently applied to large-scale agricultural datasets, where its flexibility in handling diverse input variables enables accurate predictions. This study highlights the application of KNN in crop yield forecasting, helping farmers select suitable agricultural practices and allocate resources efficiently across different climatic regions. Such predictive models can transform agricultural systems by promoting sustainable methods and contributing to long-term food security.

The strength of KNN lies in its ability to measure the similarity between new and historical data. By considering features such as soil type, rainfall, temperature, and crop characteristics, the algorithm generates accurate yield predictions that assist farmers in making key decisions about planting, irrigation, and resource distribution. Prior research has also shown that KNN and other ML models reduce the risk of crop failures and can be retrained to suit specific

ICITSM-Part II 2025, April 28-29, Tiruchengode, India
Copyright © 2025 EAI
DOI 10.4108/eai.28-4-2025.2358037

regional conditions, offering a degree of adaptability that traditional forecasting methods lack.

In the context of climate change and a growing global population, predictive modelling in agriculture is increasingly vital. Techniques such as KNN not only improve yield prediction but also support effective resource management. By maximizing productivity while minimizing environmental strain, KNN contributes to building resilient and sustainable agricultural systems.

2 System Architecture

The proposed crop yield prediction model utilizes datasets in comma-separated value (CSV) format obtained from official government sources. These datasets include essential agricultural features such as temperature, pH levels, rainfall, and nitrogen levels, all of which influence crop growth and yield. Presenting the data in a structured format ensures suitability for machine learning analysis and model training. By relying on publicly available datasets, the system is based on up-to-date and accurate parameters that reflect actual agricultural conditions.

The training dataset forms a critical component of the architecture. It enables the model to identify patterns and relationships between input features (e.g., soil quality, weather) and the output feature (crop yield). Through this process, the model learns how environmental variables affect yield and applies that understanding to unseen data. To achieve reliable predictions, data preparation steps such as cleaning, handling missing values, and standardizing formats are applied. These procedures remove inconsistencies and noise, producing a robust dataset that improves model performance.

Information pre-processing further refines the data for analysis. Standardization ensures that all features share the same scale, preventing large-scale variables from dominating the learning process. The system is configured to process CSV input efficiently, supporting accurate interpretation by the KNN model. After pre-processing, the data is merged and normalized, allowing the model to provide reliable yield predictions that support agricultural planning and decision-making.

Once prepared, the dataset is integrated with the KNN algorithm. The model identifies the closest neighbours from the training data and uses their similarity to predict crop yields for new inputs. The architecture is designed to handle large datasets and can be extended to incorporate additional features or data sources. The final outcome is a refined predictive model that delivers actionable insights for crop management, farming practices, and resource utilization.

3 Existing System

Machine learning is already being used in agriculture for crop yield prediction. Models like Random Forest, Gradient Boosting, k-Nearest Neighbour (kNN), and deep learning networks are trained on historical yield data, soil characteristics, weather records, and satellite imagery [1], [2].

An example is YieldNet, which applies deep learning to predict maize and soybean yields using remote sensing data [3]. Other studies propose hybrid frameworks combining feature selection and optimized ML models to enhance predictive accuracy [1], [4]. Recent works also demonstrate improvements through meta-features extracted from kNN regressors [4] and hybrid CNN–RNN frameworks for capturing spatiotemporal dependencies [5], as well as GNN–RNN approaches to exploit geospatial and temporal data [6]. Overall, climate-smart agriculture tools combine weather and soil inputs with ML models to forecast yield outcomes for farmers [7].

Another category is precision agriculture platforms, which rely on IoT sensors, drones, and satellite imagery to collect field-level data. This information feeds into ML systems that optimize irrigation schedules, fertilizer application, and pest management [2]. A well-known example is the IBM Watson Decision Platform for Agriculture, which integrates AI with IoT and weather forecasts to deliver insights [8]. Likewise, John Deere's See & Spray technology uses computer vision and ML to detect weeds and spray herbicides only where needed, reducing costs and chemical use [9].

Machine learning is also central to crop disease and pest detection. Image recognition models, often convolutional neural networks (CNNs), are trained on leaf images to identify diseases at early stages [5]. The PlantVillage Nuru app is a practical example, allowing farmers to use their smartphones to diagnose crop diseases [10]. Open-source TensorFlow-based models are also widely adopted for building real-time diagnostic systems [6].

A related system is soil and weather analytics, where predictive ML models integrate soil nutrient levels, moisture data, and weather forecasts to recommend farming practices. One example is the Microsoft AI Sowing App in India, which advises farmers on the best sowing times based on predictive weather and soil data. Other platforms, like AgroAPI services, provide soil and climate analytics to support farm planning decisions [7].

Finally, there are farm management decision support systems, which serve as AI-powered dashboards. These platforms integrate multiple data sources—satellite imagery, soil and weather records, and market trends—into actionable insights. For example, CropIn Smart Farm helps farmers track farm activities and predict risks, while Agrivi uses data analytics and ML to support planning, optimize yields, and forecast profitability [7]. The Fig. 1 shows System Architecture.

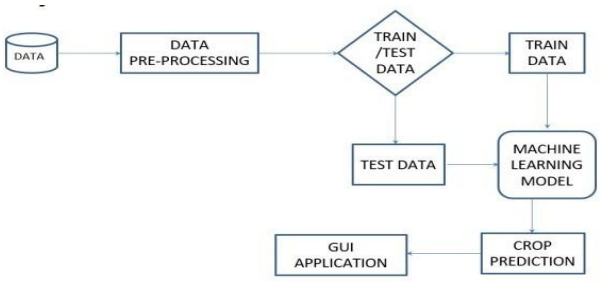

Fig. 1. System Architecture.

4 System Modules

4.1 Data Preprocessing

Data Pre-processing is an important part of raw data preparation before using them in machine learning algorithms. First clean the data by removing missing values, which might cause by problems in data acquisition or transmission. Missing data points can be removed, or the missing values can be imputed using a method like mean, median, or mode imputation. It is necessary to erase duplicated entries to eliminate redundant information which may possibly bias the results. After cleaning the data, features can be normalized/scaled as needed to get the input features on the same scale. This can be especially helpful for algorithms such as K-Nearest Neighbours (KNN), which are based model the distances between data points, as features with larger scales can unduly affect the distance calculations. Fig 2 shows the Data Collection.

District_Na	Soil_color	Nitrogen	Phosphoru	Potassium	pH	Rainfall	Temperatu	Crop	Fertilizer	Link
solapur	Black	75	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	80	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	85	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	90	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	95	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	100	50	100	6.5	1000	20	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	75	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	80	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	85	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	90	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	95	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	100	55	105	7	1100	25	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	75	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	80	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	85	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	90	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	95	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	100	60	110	7.5	1200	30	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	75	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	80	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	85	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	90	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	95	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	100	50	115	6.5	1300	35	Sugarcane	Urea	https://youtu.be/2t5Am0xLTOo
Kolhapur	Black	75	55	100	7	1400	20	Sugarcane	DAP	https://youtu.be/2t5Am0xLTOo

Fig. 2. Data Collection.

Intuitively, beyond scaled numerical features, categorical representations, such as type of crop or soil condition, need to be transformed to something that can be understood by the model. One of these popular approaches is one-hot encoding, which converts categorical features into columns of 1s and 0s which can work with many machine learning algorithms. Another important part of pre-processing is to treat outliers since they maybe scan the performance of the model. Outliers can be identified with statistical tests or graphical plots and may be subsequently removed or adjusted accordingly with respect to their contribution to the entire data distribution. By prepossessing the data, the model can work with high quality and structured data which can allow it to learn patterns from the dataset more effectively.

4.2 Feature Selection

Feature selection refers to the technique of retaining relevant features and eliminating irrelevant or redundant features for the model. For crop yield prediction, this might be finding out which features, such as temperature, rainfall, soil pH, and nitrogen content, are most important for predicting yield. One of the frequent methods employed for feature selection is the analysis of correlation, where features highly correlated with the target variable (crop yield) are determined. Those features with low correlation or redundancy are commonly eliminated to ease the model and avoid overfitting, and thus computational feasibility and generalization have been enhanced.

More robust techniques like Recursive Feature Elimination (RFE) and tree-based algorithms like Random Forests can also be utilized for this task. These methods order features by influence on the predictive power of the model. In prediction of crop yield, feature selection plays a role in to make the model capture the most important variables contributing to building the model and increasing the accuracy of the model while application. By discarding irrelevant factors, the model is more interpretable and generalizes better on unseen data because it does not overfit to noise or irrelevant patterns in the dataset.

4.3 Split Data

When the data has been processed and features have been selected, the dataset is divided into two separate collections: the training set and the test set. This separation enables it to be trained on the training set and tested on the test set, in a way that simulates what the performance of the model would be like in new, unseen data. Typically, a ratio of 80% data used for training and 20% for testing is used (though not necessarily; this may depend on the size of the data). The training data are used to train the model to learn (predict) the mapping between the input features (e.g., soil conditions and weather) and the target variable (e.g., crop yield), while the test data are used to assess the model's generalization to new data.

The breaking down of the data helps to prevent model overfit/underfit, because both training and validation data are fed into the model. Overfitting happens when a model gets too good on the training set, but its performance on new data deteriorates as the model is only memorizing the training data rather than learning generalizable pattern. Various performance metrics including MAE, RMSE or R-squared can be calculated to test the model on the test set and measure how good the model is in predicting crop yields. This step is essential for evaluating the practical utility of the model in the agricultural decision-making system.

4.4 KNN Model Training

The K-Nearest Neighbours (KNN) algorithm is a well-known and widely used simple machine learning technique for both classification and regression problems. The K Nearest Neighbours (KNN) algorithm is used for crop yield prediction where it determines the K-Nearest data in the training set for a given.

test case and predicted its yield using the average of those neighbours' yields. In KNN, the principle hyperparameter is K which is the number of neighbours used in prediction. A small

K has the property that the model will be sensitive to noise, but the predictions will be noisier One the other hand, with a larger K the predictions will be smoother as the average would be computed over more points. Choosing an appropriate value for K is key to striking a balance between model complexity and prediction accuracy.

The KNN approach which is employed in this work, does not impose any a priori assumption about true data distribution, it is very flexible. This property makes it appropriate for crop yield prediction, where the relationship between yield and environmental condition may not be linear. KNN works especially well with multi-dimensional data as it considers multiple variables all at once such as rain, temperature, soil pH and so on. Fig 3 & 4 shows the Source Code. The trained KNN model could then be used to predict crop yields for new data points by measuring the distance between the new data points and the closest neighbouring data points in the training set, in a data-driven and highly robust manner to predict crop yields.

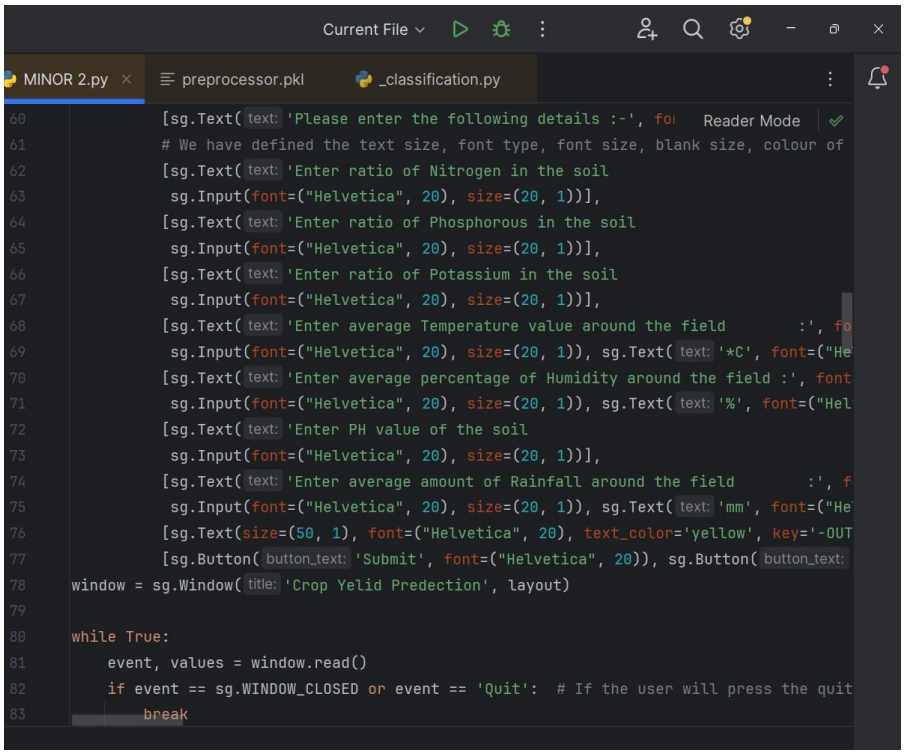

Fig. 3. Source Code.

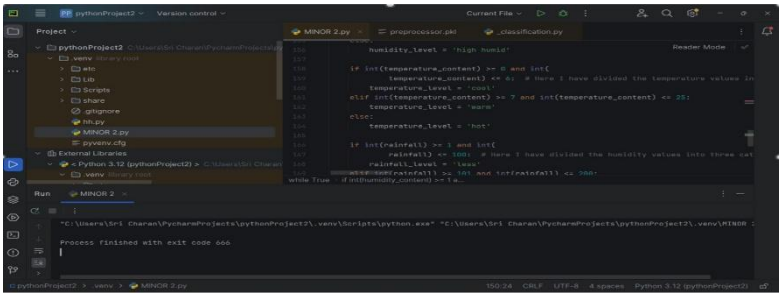

Fig. 4. Source Code.

4.5 Prediction

Once you have trained the KNN model, you can then make predictions with it on new, unseen data points (x_test) such as those from the test set. Prediction is performed by measuring distances of the test instance from training instances and predicting the crop yield by the average yield obtained from K nearest neighbours. The performance of the model can be evaluated by computing several evaluation metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared, which reflects the accuracy and stability of predictions. Fig 5 shows the Prediction Graph. These metrics can be used to compare between models and to help develop an understanding of how well the model is generalizing to new data.

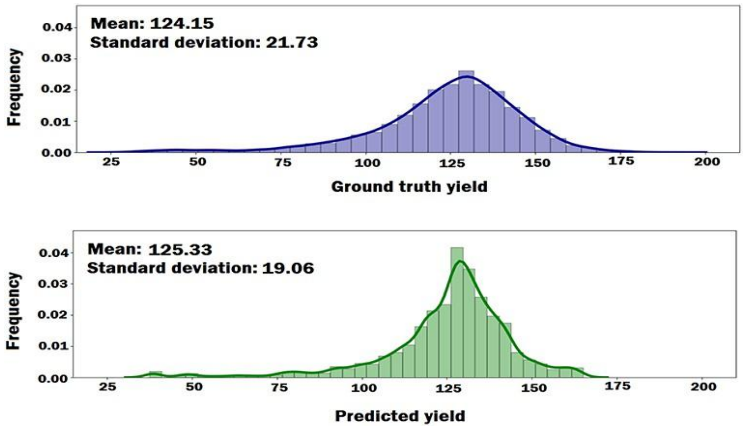

Fig. 5. Prediction Graph.

The final objective of the application of the KNN model for the prediction of crop yield is the supply of accurate forecasting to the farmers, which can support them in making decisions with regard to crop management, irrigation, and harvesting. Farmers can make maximum use of the available resources and mitigate crop failure through forecasting crop yield utilizing

relevant historical and environmental data. With continued testing and some tuning over time, the model can become more accurate and provide useful insights into crop performance under different situations. Being able to make an accurate prediction is of crucial importance for farmers in order to take data-informed decisions and hence adopt sustainable agriculture.

5 Results and Discussion

The K-nearest neighbours (KNN) model was applied to predict crop yield after considering environmental factors like temperature, rainfall, soil pH, and nitrogen content. A finely pre-processed dataset was used to train the model and its performance was then evaluated with the RMSE metric. RMSE measures the magnitude of the error (average error) but a lower score indicated better prediction accuracy. A RMSE value of 3.5 for instance, means the model, on average, is being off by 3.5 units from realizing the yield. This measure makes it possible to control the performance of the model and the places it needs to be enhanced.

One of the most import factors affecting the performance of KNN models is the choice of the value of hyperparameter K, which indicates the number of nearest neighbours to be used for prediction. If the value of K is small, then the model can lead to overfitting, becoming overly sensitive to noise in the training data and does not generalize well on new, unseen data. Or conversely, if K is too large, it may cause underfitting where the model is too simple and does not capture important patterns in the data. In order to obtain the optimal behaviour, one must choose a good K (through experimentation or methods like cross-validation etc). Cross-validation allows to select the K that minimizes the RMSE and verify that model generalizes well to different sets of data. Fig 6 shows the Output.

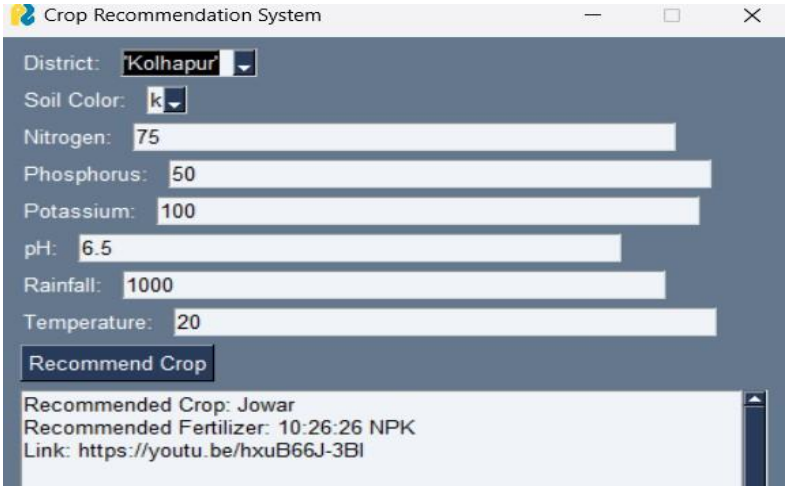

Fig. 6. Output.

The non- parametric nature of KNN makes it a particularly at- tractive candidate in crop yield prediction. In contrast to parametric models, which require a priori assumptions on the functional relationship between the input features and the target value, KNN makes no assumption regarding the underlying data distribution. This flexibility enables KNN to adjust to different data patterns and makes it applicable in complex agricultural situations where the relationship between environmental factors and crop yield can be nonlinear. However, even though KNN is flexible, it may pose some problems, in particular when dealing with large amounts of data, like the need to compute distances between the test point and all training points in at the prediction step. The computational burden scales with dataset size and larger datasets can result in longer prediction times and higher memory requirements.

KNN algorithm is also sensitive to feature scaling. So, as KNN computes the distance between data points, since larger numerical values tend to overshadow the effect of small scaled features, scaling is done so that the effect of each feature can be equally considered. The data should then be standardized or normalized before modelling. Without scaling, the KNN model may weigh some features too much, it could affect the predictions. Feature scale methods, including Min-Max scaling or Z-score normalization, can be used to align all variables to same scale such that all features contribute the same in calculations of distance and improves the performance of the model.

In summary, the KNN model has potential to predict crop yield using environmental features, but the model is affected by several main factors: \(K\), feature scaling and computation costs. With the correct choice of K value and handle issues related to feature scaling properly, you now have a great model i.e. KNN model for agriculture yield prediction. Such an efficiency may deteriorate as the dataset gets larger, calling for an efficient data processing as well as an adoption of more scalable algorithms for large agricultural datasets. Future work may consider more complex models or additional data sources like satellite images or real-time sensor measurements to further improve the accuracy of crop yield predictions.

6 Conclusion

This work demonstrates the effectiveness of the K-Nearest Neighbours (KNN) algorithm in predicting crop yield using environmental variables such as temperature, rainfall, soil pH, and nitrogen levels. The model, evaluated with Root Mean Squared Error (RMSE), shows reliable performance, though its accuracy is highly dependent on the choice of the hyperparameter

K

K. Fine-tuning

K

K through cross-validation is essential to ensure good generalization and to avoid both overfitting and underfitting.

Despite its advantages, KNN has limitations when applied to large datasets, as distance calculations for all training samples increase computational time and resource requirements.

These challenges are particularly significant in real-time applications, where rapid predictions are critical for agricultural decision-making. Furthermore, since KNN relies on distance measures, appropriate feature scaling is necessary to prevent certain variables from dominating the model and to ensure balanced, unbiased predictions.

In summary, KNN offers strong potential for crop yield prediction, provided that key considerations such as hyperparameter selection, feature scaling, and computational efficiency are carefully addressed. Future research may incorporate additional data sources, including satellite imagery and sensor-based inputs, to enhance prediction accuracy. Exploring advanced machine learning approaches, such as ensemble models and deep learning techniques, could further improve scalability and robustness, supporting more accurate and sustainable agricultural forecasting.

References

[1] Shawon, S. M., Barua Ema, F., Mahi, A. K., & Mohsin Sarker Raihan, M. (2023). Crop yield prediction: Robust machine learning approaches for precision agriculture. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/ICCIT60459.2023.10441634

[2] Barbosa, L. A. F., Guilherme, I. R., Pedronette, D. C. G., & Tisseyre, B. (2025). Meta-Features Extracted from Use of kNN Regressor to Improve Sugarcane Crop Yield Prediction. *Remote Sensing*, *17*(11), 1846. https://doi.org/10.3390/rs17111846

[3] Abdel-Salam, M., Kumar, N., & Mahajan, S. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. *Neural Computing & Applications*, *36*(1), 20723–20750. https://doi.org/10.1007/s00521-024-10226-x

[4] Yan, Y., Wang, Y., Li, J., Zhang, J., & Mo, X. (2025). Crop yield time-series data prediction based on multiple hybrid machine learning models. *arXiv*. https://doi.org/10.48550/arXiv.2502.10405

[5] Shawon, S. M., Ema, F. B., Mahi, A. K., Niha, F. L., & Zubair, H. T. (2025). Crop yield prediction using machine learning: An extensive and systematic literature review. *Smart Agricultural Technology*, *10*, 100718. https://doi.org/10.1016/j.atech.2024.100718

[6] Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 1750. https://doi.org/10.3389/fpls.2019.01750

[7] Fan, J., Bai, J., Lin, Z., Ortiz-Bobea, A., & Gomes, C. P. (2022). A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. *arXiv*. https://doi.org/10.48550/arXiv.2111.08900

[8] Nosratabadi, S., Imre, F., Szell, K., Ardabili, S., Beszedes, B., & Mosavi, A. (2020). Hybrid machine learning models for crop yield prediction. *arXiv*. https://doi.org/10.48550/arXiv.2005.04155

[9] Sikarwar, S. S., Pandey, S., Kumar, S. A., Kumar, P., Sahu, P. K., & Singh, B. (2024). Optimizing crop yield predictions using K-nearest neighbors regression: An analysis of temperature, rainfall and soil pH influences. *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, 1336–1341. https://doi.org/10.1109/ICAC2N63387.2024.10894947

[10] Latha Jothi, V., Neelambigai, A., Nithish Sabari, S., & Santhosh, K. (2020). Crop yield prediction using KNN model. *International Journal of Engineering Research & Technology (IJERT): RTICCT – 2020 Conference Proceedings*, 2278-0181.