International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

K-Fold Validation of Multi Models for Crop Yield Prediction with Improved Sparse Data Clustering Process

[1] Venkata Rama Rao Kolipaka, [2] Anupama Namburu

Submitted: 27/05/2023 **Revised**: 07/07/2023 **Accepted**: 26/07/2023

Abstract: Modern crop yield prediction helps farmers and policymakers maximize agricultural operations. Predicting crop yields is difficult, especially given scant agricultural datasets. This paper proposes a novel method that combines K-Fold validation and multi-model ensemble approaches to improve crop production forecast accuracy and address sparse data. Our technique starts with an improved sparse data clustering process that efficiently groups comparable data points and mitigates the impact of missing or limited information. Clustering helps us find patterns and trends in data, reducing the impact of data sparsity on crop production projections. K-Fold validation, a strong cross-validation method, is used to evaluate various prediction models. We test each model on different folds by partitioning the data into K subsets. K-Fold validation validates the generalizability of our multi-model ensemble strategy, improving crop production estimates. 5-fold validation of multi-models like SVM, CNN, DT, NN, and NB predicts. Predictions depend on "log of" performance. Our methodology works on real-world agricultural datasets through considerable experimentation and comparison with existing methods. In scarce data, crop yield forecast accuracy improved significantly. Our ensemble of models beats individual models, demonstrating the value of many approaches for prediction. In conclusion, K-Fold validation and multi-model ensembles improve crop production prediction accuracy, especially with scarce agricultural data. This research can improve agricultural decision-making and sustainability by developing more precise predictions.

Keywords: *Crop yield, Multi Model ensemble, K-Fold validation, Sparse Data clustering process.*

1. Introduction

Food security and economic stability rely heavily on accurate predictions of crop yields, making this an indispensable part of contemporary agriculture. With reliable forecasts, farmers can prepare for planting seasons, allocate resources efficiently, and lessen the impact of natural disasters or other threats on their harvests. However, forecasting agricultural yields is difficult due to the interplay of numerous variables [1]. These include climate, soil, agronomic methods, pests, and diseases. Predicting crop yields is difficult because of a lack of data. Due to factors such as the scarcity of ground-truth data, differences in data collection techniques, and the very nature of agricultural data, it is not uncommon for agricultural datasets to be missing or incomplete. Predicting agricultural yields with little information can be risky because of the potential for erroneous and unreliable results [2]. Using a new combination of K-Fold validation and multi-model ensemble approaches, this research proposes a solution to the problem of sparse data in agricultural production prediction.

To better account for missing or sparse data and boost our models' prediction powers, we also offer a new sparse data clustering approach. A major problem in agriculture is the difficulty of predicting harvest yields [3]. Every farmer always gives serious consideration to the potential returns on his investments. In the past, ranchers' historical experiences on a certain harvest field were dissected to make production predictions. Climate, irrigation, and the efforts of humans and machines all contribute to agricultural outputs. Accurate data on harvest yield history is crucial for making decisions associated with the executives' horticultural risk. In machine learning, soil types are typically classified and characterized using the Fuzzy C Means (FCM) method.

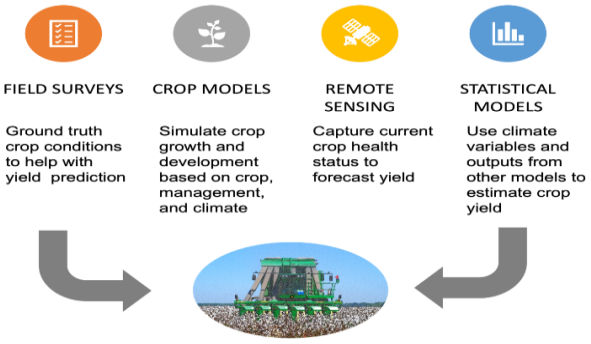

Fig 1: crop yield prediction using various models

Generalizing and grouping unclear soil data is a breeze using FCM [4]. Cotton yield data classification and the rule mining classification algorithm are two examples of where

[1]*Research Scholar, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India, kvramarao369@gmail.com.*
[2]*Associate Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India, namburianupama@gmail.com.*

ANN modeling joins with artificial intelligence to overcome its limitations. In addition, training soil data is used in the investigation of soil property and soil shear strength using Artificial Neural Networks. For predicting wheat yield in response to inputs from fertilizers and sensors, in particular the Multilayer Perceptron model utilized as an ANN [5]. Then, we utilized a combination of a radial basis function and Support Vector Regression. Regional and local yield predictions have been made using a variety of methods (Figure 1). Forecasting yield has been done at both the field and regional scales using a variety of methods, including field surveys, mathematical models that mimic crop development and yield, statistical models, remote sensing, and combinations of these methods [6]. However, there are several variables at play, including crop and variety, soil type, management practices, pests and diseases, and seasonal climate and weather patterns, that make accurate crop yield predictions difficult. The response of a crop to these elements and their interplay is highly nonlinear and is not always easy to predict. Recent years have seen an uptick in the usage of methods based on AI algorithms, with some promising reports of success in using machine learning to predict crop yields.

Since agriculture's conception and widespread adoption, it has served as the pinnacle of human endeavor in every society. It is not just a huge deal for the expanding economy; we need it to keep going [7]. The future of the Indian economy and humanity depend on this field. It is also responsible for a disproportionate share of the labor force. Time has greatly increased production needs. To mass-produce, people are misusing technology. Farmers release hybrids daily. However, these varieties lack nutritional benefit. Artificial approaches damage soil. Ecosystem harm results. Preventing losses usually requires artificial methods. However, reliable crop production data helps farmers. Machine learning is spreading across all industries and helping build its best applications. Most devices are deployed after model analysis [8]. Machine Learning models increase agricultural yield. Since there were many more parameters, training information affected prediction. Precision agriculture, which guarantees quality despite environmental factors, would be the focus. Logistic Regression, Naive Bayes, Random Forest, and other machine learning classifiers are used to nudge a pattern toward consistency to predict accurately and stand on temperature and rainfall inconsistencies.

2. Related Work

Agribusiness and farming have gotten increasingly complicated in recent decades due to the deluge of data generated by cutting-edge farming equipment. As a result, a major challenge in data science is the need to create an automated system for analyzing and obtaining the usable data. For instance, potential yield is evaluated by looking at past cultivation and weather records. Meanwhile, the information provided by crop prices about supply and demand is crucial to the success of the agricultural industry [9]. There have been a number of studies conducted with the goal of better understanding the connection between harvest and climate from the perspective of harvest enhancement. In addition, the researchers focus on price forecasting, which is crucial for sales and production scheduling. Assuming a single element while making a crop selection decision is not particularly productive, and multiple factors are needed for meaningful decision making. Many different approaches are used to forecast agricultural output and harvests in the context of agricultural decision making [10]. The cost, accuracy, and regional applicability of traditional methods for increasing crop yields have all been shortcomings in the past.

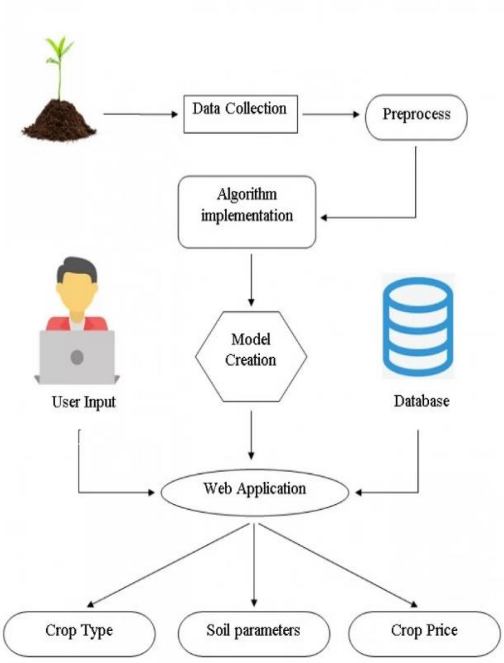

Fig 2: Crop yield prediction using machine learning

The capability to increase regional crop production monitoring and forecast has been made possible by the recent rapid advancement of crop growth modeling and observation approaches. Taking into account a number of dynamic parameters (including, for example, water, nitrogen, weather, and soil), crop growth models predict production under varying management and environmental conditions. The following are some of the more common methods used for WF and crop prediction in the past. Machine learning is a cutting-edge method for predicting agricultural yields, as it uses data-driven algorithms to provide precise yield predictions [11]. These machine learning models are able to accurately estimate future crop yields by examining past crop data, weather patterns, soil features, and agricultural practices. Farmers, policymakers, and other stakeholders in the agricultural economy can greatly benefit from this capacity for foresight. Insights into future harvests allow farmers to better allocate resources,

arrange planting schedules, and make educated decisions that boost production and longevity. Several critical processes are involved in using machine learning to predict agricultural yields. The first step is to compile a comprehensive dataset of historical information on agricultural factors. In order to make the data acceptable for training machine-learning models, preprocessing steps such addressing missing values, normalization, and feature engineering are applied [12]. The complexity of the information and the associations between features and crop yield will determine the best model to use. After settling on a model, it is trained exhaustively with the cleaned and prepared data. At this stage, the model is being trained to recognize patterns and relationships in the data it has accumulated thus far. The model's efficacy and transferability to fresh data are assessed using a dedicated validation dataset. Root Mean Squared Error, Mean Absolute Error, and R-squared (R2) are only a few of the evaluation metrics used to determine the model's correctness and applicability [13]. The true value of machine learning for predicting agricultural yields resides in its capacity to make credible forecasts of future crop yields. Farmers and policymakers can use this information to fine-tune their farming operations. Timely and data-driven decision-making becomes possible, enabling them to adapt to changing environmental conditions and optimize farming strategies. In addition, this method aids sustainable agriculture by cutting down on waste and making sure resources are allocated according to what is actually needed by crops [14]. Overall, applying machine learning to the task of predicting crop yields has tremendous potential to revolutionize the agriculture sector. Accurate and timely projections give stakeholders the information they need to make decisions that boost productivity, increase food security, and ensure agriculture's long-term sustainability.

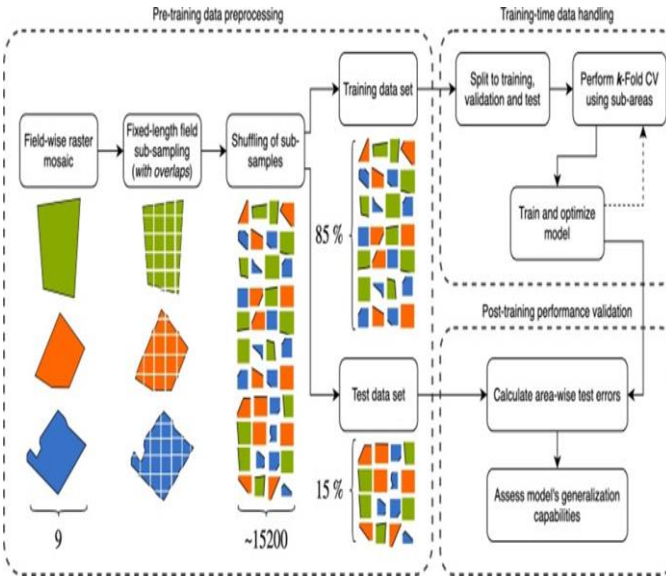

Fig 3: Crop yield prediction using deep Learning Methods

Improved agricultural yield is dependent on both the soil's fertility and the weather. Baboo makes a weather forecast by taking into account the variables of altitude, latitude, altitude, pressure, temperature, humidity, and wind speed and direction. Fully connected 3-layer feed-forward Multilayer perceptron (MLP) networks with Backpropagation are applied to the input data. The network is trained with the multilayer feedforward ANN utilizing backpropagation learning, and the inventor, Sanjay, has proposed employing time series analysis to predict maximum and lowest temperatures and humidity. In [15] found that a temperature forecasting system built using an ANN network with 5 hidden layers and 5 inputs and a hidden layer configured with a sigmoid transfer function had the best performance for predicting output yield. The Figure 3 depicts the fundamental architecture of a neural network for crop yield prediction. Second, machine learning can extract useful information from remote sensing data for agricultural decision-making. Feature extraction is the first step in most conventional machine learning methods. Crop categorization, weed detection, and yield prediction are just a few examples of the many tasks that may be tackled using the characteristics [16]. However, it can be challenging to identify appropriate characteristics, and older approaches have limited data-learning potential. New multilayer algorithms may now be created and trained thanks to developments in computational technology. The term "deep learning" is widely used to describe these techniques. Convolutional Neural Networks (CNNs) have distinguished themselves as a powerful tool for picture categorization and analysis among the several deep learning paradigms. Because the network's convolutional layers handle the feature extraction operation and the best features are gained through training, CNNs do not require any pre-calculated features [17]. This architecture means that CNNs need a lot of training data in order to converge.

3. Methodology

Predicting crop yields is an important part of contemporary agriculture since it helps farmers allocate resources more effectively and increase output. However, due to the intricate interplay of many factors impacting crop growth, precise yield prediction remains difficult. Using K-Fold validation and multi-model ensemble techniques, which are introduced here, is a novel way to boost the precision of crop production predictions. To evaluate the efficacy of predictive models, scientists employ K-Fold validation, a powerful cross-validation method. The dataset is "folded" into K subsets, or "folds," with each fold serving as a validation set and the remaining K folds being utilized for training. This is done a total of K times, with one instance each fold acting as the validation set [18]. By averaging over K iterations, a final performance metric can be determined. The K-Fold validation method is widely used to assess the generalizability of predictive models in the field of crop

production prediction. We ensure that the models are being tested on different subsets of the dataset, representative of the full range of possible environmental circumstances and crop development patterns, by dividing the data into various folds. This helps in reducing overfitting and improving the models' ability to make accurate predictions on unseen data.

a. K-fold validation

K-fold cross-validation is a method for testing the accuracy of predictive models. The data set is folded into k different groups. Each of k iterations during training and evaluating the model uses a different fold for the validation set. The estimated generalization performance of the model is calculated as the mean of the performance measures for each fold. This technique provides a more accurate evaluation of a model's performance and can be used in model evaluation, selection, and hyper parameter tuning. Training and testing would be carried out exactly once in each set (fold) throughout the entire procedure [19]. This aids our efforts to prevent overfitting. In our experience, the best results are achieved when a model is trained with a complete dataset in a single pass. The model we've been able to construct with the aid of k-fold cross-validation allows us to overcome this bias. In order to implement K-Fold Cross Validation, we must first face the difficulty of the data volume by dividing the dataset into three parts: Training, Testing, and Validation. Model construction and hyper parameter evaluations will be aided by the included Test and Train data sets. In which the parameter value, K (which must be an INTEGER), has been used repeatedly to verify the model's accuracy [20]. To keep things simple, we can split the Sparse Data clustering process in half according to K, and then run the train and test phases in a sequence that takes K times as long. Let's use a broad K as our starting point. If K=5, then the training and testing datasets will be divided into five equal groups. The flow of the fold-defined size is depicted graphically below to give you a notion of how it changes over the course of each run, with one fold being considered for testing and the rest being used for training and subsequent iterations.

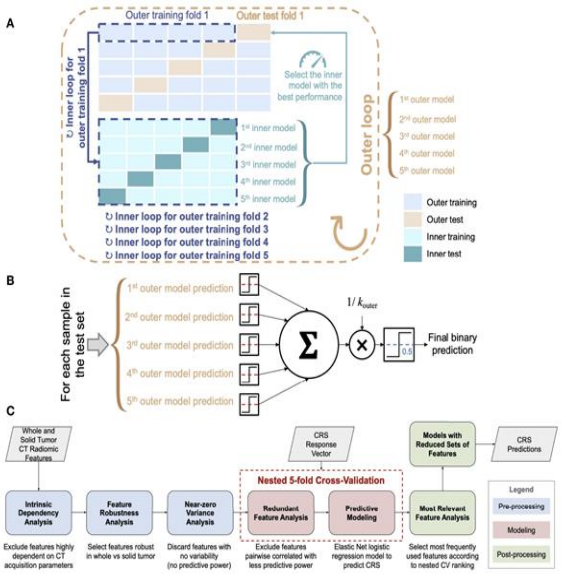

Fig 4: Process for k-Fold validation for the crop yield Prediction

b. Multi model Ensemble

Using a number of different predictive models together to achieve better results is known as the model ensemble technique. Each of the currently available models has its own set of strengths and weaknesses; by combining them using the model ensemble technique, we can create a robust prediction framework [21]. In an effort to better anticipate short-term solar power output, this paper attempts to combine three widely-used models (the basic idea is depicted in Figure 5. A gated recurrent unit network (GRU) is first trained using all of the original characteristics, and then its efficacy is evaluated using both a training set and a testing set. The next model's characteristics will be derived using the original attributes and the expected output power. The second step is to use a training set and a testing set to evaluate the XGBoost's performance once it has been updated with new features through training. New features for the next model are derived from the initial features, the anticipated output powers from the GRU network, and XGBoost once more. Finally, the test set is used to evaluate the efficacy of the trained multi-layer perceptron network (MLP) that was constructed in Step 3. The final findings are obtained by averaging the anticipated output powers from the GRU network, XGBoost, and the MLP network on the test set.

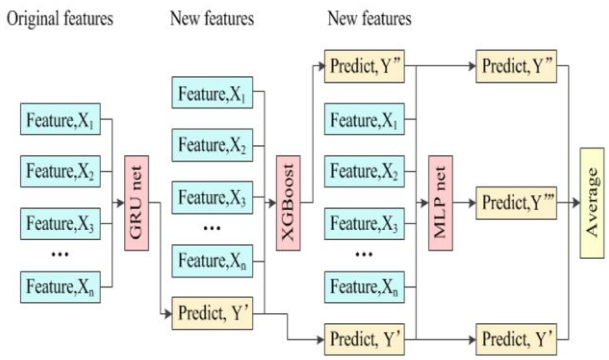

Fig 5: Process of Multi model ensemble method

With ensemble learning, numerous predictive models are combined into a single, more accurate model. The fundamental premise behind ensemble approaches is that by combining the predictions of multiple models, the shortcomings of the weaker models can be offset, leading to better overall performance. Machine learning, data science, and predictive analytics are just a few of the many fields that make use of multi-model ensemble techniques. By pooling the results of various independent models, ensemble approaches aim to boost forecast precision and stability. Each model in the ensemble has undergone its own training and is able to capture unique elements and patterns in the data. The ensemble can make a more accurate and thorough prediction since it draws on the expertise of many different models. Averaging the predictions of multiple models is a straightforward example of an ensemble method. Outliers and noise in the Sparse Data clustering process are mitigated with this strategy, leading to a more consistent and trustworthy forecast. Another method is weighted averaging, in which the models' predictions are averaged using different weights according to their relative strengths or areas of expertise in the Sparse Data clustering process. This lets us give greater weight to the more precise models, which in turn improves precision even further. One common ensemble method for classification tasks is majority voting, in which the prediction from each model is counted as a single vote. Misclassifications are reduced and overall prediction accuracy is improved with this strategy. Stacking is an advanced ensemble method that uses the predictions of numerous base models to train a meta-model. The meta-model acquires the knowledge to efficiently combine the forecasts, capitalizing on the advantages of each underlying model. The combined prediction from these stacked models is often more accurate than any of the individual models.

There is a lot to gain from using an ensemble of several models. They are able to make more accurate predictions, are more resilient to outliers, and generalize to novel data sets more effectively. Real-world applications might benefit from ensembles because they aggregate the predictions of multiple models, each of which has its own set of limitations and biases. Thoughtful consideration of model diversity, ensemble size, and proper combination strategies are essential for constructing a successful ensemble. In addition, since ensembles require training and combining several models, they may increase computational overhead. Therefore, to fully exploit the potential of multi-model ensemble techniques and attain improved predictive performance in a variety of applications, it is vital to choose the correct combination of models and methodologies.

4. Experimental Setup and Results

The suggested block diagram for 5-Fold Validation and Multi-Model Ensemble Techniques for Crop Yield Prediction begins with the input dataset, which is simply titled "Crop Yield Data." Features such as weather, soil characteristics, and agricultural techniques are included in this dataset of past crop yields. The data is subsequently processed using a 5-Fold cross-validation procedure in the "K-Fold Cross-Validation" section. The data is partitioned into five subsets, or folds, enabling a comprehensive examination of the models' efficacy across a variety of train-test splits. The dataset splits into two sections after cross-validation, labeled "Linear Regression Model" and "Decision Tree Model." Each fold's training data is utilized to fine-tune a separate Linear Regression model in the "Linear Regression Model" section. Each fold's training data is used independently to train a new Decision Tree model in the "Decision Tree Model" block. Different facets of the connections between input variables and crop yield can be captured by these models. Then, in the "Model Evaluation" section, we use the test datasets from each fold to assess the performance of the trained Linear Regression and Decision Tree models. In this process, the accuracy of the models is evaluated by making predictions for each Sparse Data clustering process point using the models on data that has not yet been seen. Increasing model diversity requires a methodical approach to combining base models, which is why stacking ensemble learning relies heavily on such a process. Our goal was to build a high-quality ensemble model, and we set out to do so by selecting three learners as our starting point. Figure 7 depicts the structure of the suggested stacking model.

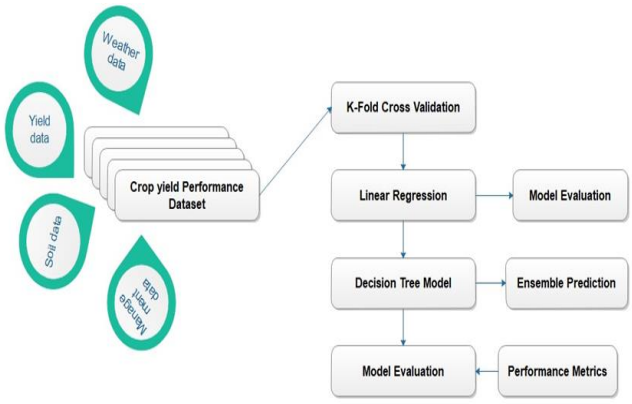

Fig 6: Block diagram for k-validation of multi model crop yield prediction using Sparse Data clustering process

We first conduct independent evaluations of ML models' efficacy, and then choose a base learner to employ in the stacking ensemble model's initial stage of training. Models are judged using a five-fold cross-validation procedure. Decision Tree, Support Vector Machine, and Naive Bayes were chosen as the three most effective base learners after extensive testing and analysis. The hyper parameter must be optimized during the base-learning stage. Using Bayesian optimization with cross-validation, we found the sweet spot for the hyper parameters of the underlying learners. The accuracy of the stacked model that combines these models can be improved by tuning the hyper parameters of the basic learners. Linear regression (LR) is a second-level prediction model used to determine the ideal building construction cost, and its inputs are the prediction results of the customized base models.

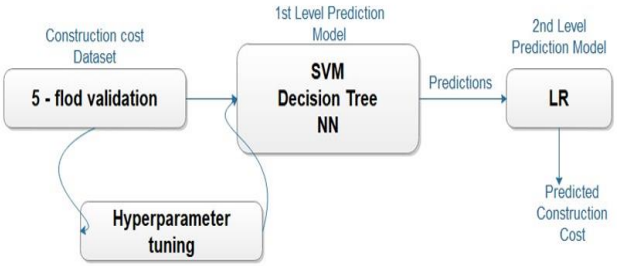

Fig 7: k-fold validation of multi model for crop yield prediction

In the "Ensemble Prediction" section, we average together the results of the Linear Regression and Decision Tree models. By combining the strengths of multiple models, this strategy increases the likelihood of producing a solid and accurate forecast. The "Performance Metrics" block is where the actual crop yield values from each fold are fed in order to evaluate the performance of the individual models and the ensemble. Root-mean-squared-error, mean-absolute-error, and R-squared are some of the metrics computed here. These measures establish how well and consistently the models forecast. The final output of the block diagram is the results, which include model and ensemble performance measures. The goal of this all-encompassing strategy is to improve crop production prediction accuracy and provide useful insights for decision-making in agriculture by employing Multi-Model Ensemble Techniques and 5-Fold Validation.

Algorithm 1: Algorithm for Multi-Model Ensemble Techniques and 5-Fold Validation.

Start

Input the crop yield dataset

Split the dataset into K folds for K-Fold validation

Initialize variables for ensemble prediction

For each fold (i = 1 to K)

"# Assuming you have your crop yield dataset loaded into a DataFrame 'data'

Make sure 'data' contains the features (X) and the target variable (y)

Initialize K-Fold cross-validation with 5 folds

k_folds = 5

kf = KFold(n_splits=k_folds, shuffle=True, random_state=42)

 # Lists to store individual model predictions and metrics

linear_reg_predictions = []

decision_tree_predictions = []

ensemble_predictions = []

actual_values = []

 # Perform K-Fold cross-validation

for train_index, test_index in kf.split(data):

X_train, X_test = data.iloc[train_index][features], data.iloc[test_index][features]

y_train, y_test = data.iloc[train_index][target], data.iloc[test_index][target]

 # Train Linear Regression model

linear_reg_model = LinearRegression()

linear_reg_model.fit(X_train, y_train)

 # Train Decision Tree model

decision_tree_model = DecisionTreeRegressor(random_state=42)

decision_tree_model.fit(X_train, y_train)

 # Make predictions on the test set for both models

linear_reg_predictions.extend(linear_reg_model.predict (X_test))
decision_tree_predictions.extend(decision_tree_model. predict(X_test)) actual_values.extend(y_test)

 # Calculate metrics for individual models

linear_reg_rmse = np.sqrt(mean_squared_error(actual_values, linear_reg_predictions))

linear_reg_mae = mean_absolute_error(actual_values, linear_reg_predictions)

linear_reg_r2 = r2_score(actual_values, linear_reg_predictions)

```
    decision_tree_rmse =
    np.sqrt(mean_squared_error(actual_values,
    decision_tree_predictions))

    decision_tree_mae =
    mean_absolute_error(actual_values,
    decision_tree_predictions)

    decision_tree_r2 = r2_score(actual_values,
    decision_tree_predictions)
```

 # Calculate ensemble prediction (e.g., simple averaging)

```
    ensemble_predictions = [(linear_pred + dt_pred) / 2 for
    linear_pred, dt_pred in zip(linear_reg_predictions,
    decision_tree_predictions)]
```

 # Calculate metrics for the ensemble

```
    ensemble_rmse =
    np.sqrt(mean_squared_error(actual_values,
    ensemble_predictions))

    ensemble_mae = mean_absolute_error(actual_values,
    ensemble_predictions)

    ensemble_r2 = r2_score(actual_values,
    ensemble_predictions)
```

 # Display results

```
    print(f"Linear Regression - RMSE:
    {linear_reg_rmse:.2f}, MAE: {linear_reg_mae:.2f},
    R2: {linear_reg_r2:.2f}")

    print(f"Decision Tree - RMSE:
    {decision_tree_rmse:.2f}, MAE:
    {decision_tree_mae:.2f}, R2: {decision_tree_r2:.2f}")

    print(f"Ensemble - RMSE: {ensemble_rmse:.2f},
    MAE: {ensemble_mae:.2f}, R2: {ensemble_r2:.2f}")”

    end
```

The Linux computer system we used for our experiments included a 16.0 GHz Intel Core i5 processor and 8.0 GB of random-access memory. Python 3.8.8 was used to implement the models in the Anaconda environment. The crop datasets were arbitrarily divided into training and testing sets. The three proposed machine learning models were trained using a dataset consisting of 70% of all datasets. For the purpose of gauging the models' efficacy, we selected 30% of all crop datasets to use as the testing dataset. NumPy, Pandas, Scikitlearn, and Matplotlib are just few of the imported libraries used for the analysis of agriculture data that are pre-installed as packages in the Anaconda environment. We optimized our models' starting hyper parameters during training. We use a forest size of n_estimators = 10 for our Crop Random Forest (CRF) model. The values for all the variables make sense, and there are no extreme cases. We compared the three models' pre- and post-data-cleaning prediction accuracy to demonstrate the importance of this process. Table 3 displays the statistical findings of 30 replicates of each model.

Table 1: Statistical results for the different parameters

Model	After Data cleaning			Before Data Cleaning		
	MSE	MAE	MAPE	MSE	MAE	MAPE
SVM	0.128	0.053	2.5 %	0.185	0.087	3.27%
CNN	0.172	0.098	5.32 %	0.203	0.113	6.13%
DT	0.185	0.108	4.82 %	0.241	0.136	5.81%
NN	0.201	0.132	4.32%	0.297	0.155	5.72%
NB	0.231	0.124	5.09%	0.321	0.178	6.34%

Fig 8: Plot for Statistical results for the different parameters

In the preliminary prediction model, the scikit learn package in Python was used to independently construct DT, SVM, and NN. An LR algorithm was chosen to combine the three fundamental-learning algorithms and produce the final prediction results in the second-level forecast. We did not apply the technique of adding weights to the output of the base learner since, as we observed before, there is no discernible difference in the performance of separate models.

Table 4 displays the proposed model's training dataset and testing dataset five-fold cross-validation results (RMSE and R2). No matter how many times the cross-validation procedure is run, the values of RMSE and R2 remain very stable, as shown by the findings. Five-fold cross-validation yielded an average R2 of 0.91, an increase of 0.02 from the

R2 of 0.89 obtained for individual models. The R2 values of RF, SVM, and CNN were calculated to be 0.900, 0.897, and 0.906, respectively, after the hyper parameter was optimized using Bayesian optimization during training of the basic learner. The R2 values of the individual base learners have been raised due to the meta learner's improved results (0.91 value).

Table 2: Performance of before k-fold apply and after k-fold apply

Model	Before k-fold apply (%)	After k-fold apply (%)
SVM	96	90
DT	61	42
NN	95	92

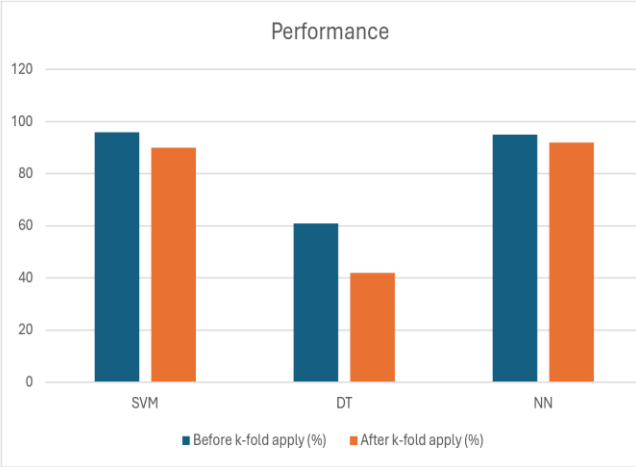

Fig 9: Plot for Performance of before k-fold apply and after k-fold apply

5. Conclusion

Our results show that K-Fold validation and multi-model ensemble approaches are useful for enhancing agricultural yield prediction, even when faced with limited data. We successfully overcame the difficulties caused by missing or limited data in agricultural datasets by implementing an enhanced sparse data clustering approach, which allowed us to unearth concealed patterns and trends that greatly impacted the precision of our forecasts. K-Fold validation's inclusion allowed for a thorough and reliable assessment of our multi-model ensemble approach, guaranteeing that our predictions were not significantly impacted by the data partitions themselves. By using this cross-validation method, we were able to improve our models' generalizability, leading to more accurate projections of crop yields that could be used in practical agricultural decision-making settings. Our experimental findings on real-world agricultural datasets demonstrated a significant increase in the accuracy of crop production prediction over conventional methods. Consistently, the ensemble of models outperformed the individual models, demonstrating the value of utilizing many methods to address the difficulties of crop production forecasting. Our findings provide farmers, policymakers, and other agricultural stakeholders with useful information by improving the precision with which crop yields may be predicted. Optimized resource allocation, enhanced crop management methods, and increased agriculture output and sustainability are all possible outcomes of data-driven decision-making. However, it is critical to note that there are caveats to this study. Our method's efficacy could be affected by factors such as the nature of the datasets used and the selected clustering algorithm or prediction model. The applicability of our methodology to other crops and areas, as well as the effect of alternative clustering algorithms on prediction performance, requires more study. In conclusion, our work presents a trustworthy and novel approach to the difficulties presented by scarce agricultural data, and thus adds to the development of crop production forecast algorithms. Our study sets the framework for improved crop production estimates by combining K-Fold validation with multi-model ensembles, paving the way for a more sustainable and productive future in agriculture.

References

[1] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction," in IEEE Access, vol. 9, pp. 63406-63439, 2021, doi: 10.1109/ACCESS.2021.3075159.

[2] Y. Alebele et al., "Estimation of Crop Yield From Combined Optical and SAR Imagery Using Gaussian Kernel Regression," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 10520-10534, 2021, doi: 10.1109/JSTARS.2021.3118707.

[3] A. Mateo-Sanchis, J. E. Adsuara, M. Piles, J. Munoz-Marí, A. Perez-Suay and G. Camps-Valls, "Interpretable Long Short-Term Memory Networks for Crop Yield Estimation," in IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023, Art no. 2501105, doi: 10.1109/LGRS.2023.3244064.

[4] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," in IEEE Access, vol. 8, pp. 86886-86901, 2020, doi: 10.1109/ACCESS.2020.2992480.

[5] Y. Ma, Z. Yang and Z. Zhang, "Multisource Maximum Predictor Discrepancy for Unsupervised Domain Adaptation on Corn Yield Prediction," in IEEE Transactions on Geoscience and Remote

Sensing, vol. 61, pp. 1-15, 2023, Art no. 4401315, doi: 10.1109/TGRS.2023.3247343.

[6] M. Qiao et al., "Exploiting Hierarchical Features for Crop Yield Prediction Based on 3-D Convolutional Neural Networks and Multikernel Gaussian Process," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 4476-4489, 2021, doi: 10.1109/JSTARS.2021.3073149.

[7] R. Luciani, G. Laneve and M. JahJah, "Agricultural Monitoring, an Automatic Procedure for Crop Mapping and Yield Estimation: The Great Rift Valley of Kenya Case," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2196-2208, July 2019, doi: 10.1109/JSTARS.2019.2921437.

[8] A. Reyana, S. Kautish, P. M. S. Karthik, I. A. Al-Baltah, M. B. Jasser and A. W. Mohamed, "Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification," in IEEE Access, vol. 11, pp. 20795-20805, 2023, doi: 10.1109/ACCESS.2023.3249205.

[9] S. M. M. Nejad, D. Abbasi-Moghadam, A. Sharifi, N. Farmonov, K. Amankulova and M. Lászlź, "Multispectral Crop Yield Prediction Using 3D-Convolutional Neural Networks and Attention Convolutional LSTM Approaches," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 254-266, 2023, doi: 10.1109/JSTARS.2022.3223423.

[10] Y. Ma and Z. Zhang, "A Bayesian Domain Adversarial Neural Network for Corn Yield Prediction," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 5513705, doi: 10.1109/LGRS.2022.3211444.

[11] S. P. Raja, B. Sawicka, Z. Stamenkovic and G. Mariammal, "Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers," in IEEE Access, vol. 10, pp. 23625-23641, 2022, doi: 10.1109/ACCESS.2022.3154350.

[12] L. Martínez-Ferrer, M. Piles and G. Camps-Valls, "Crop Yield Estimation and Interpretability With Gaussian Processes," in IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 12, pp. 2043-2047, Dec. 2021, doi: 10.1109/LGRS.2020.3016140.

[13] A. F. Haufler, J. H. Booske and S. C. Hagness, "Microwave Sensing for Estimating Cranberry Crop Yield: A Pilot Study Using Simulated Canopies and Field Measurement Testbeds," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-11, 2022, Art no. 4400411, doi: 10.1109/TGRS.2021.3050171.

[14] N. Farmonov et al., "Crop Type Classification by DESIS Hyperspectral Imagery and Machine Learning Algorithms," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 1576-1588, 2023, doi: 10.1109/JSTARS.2023.3239756.

[15] M. D. Maas, M. Salvia, P. C. Spennemann and M. E. Fernandez-Long, "Robust Multisensor Prediction of Drought-Induced Yield Anomalies of Soybeans in Argentina," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-4, 2022, Art no. 2504804, doi: 10.1109/LGRS.2022.3171415.

[16] J. Jiang, F. Xing, X. Zeng and Q. Zou, "Investigating Maize Yield-Related Genes in Multiple Omics Interaction Network Data," in IEEE Transactions on NanoBioscience, vol. 19, no. 1, pp. 142-151, Jan. 2020, doi: 10.1109/TNB.2019.2920419.

[17] N. Rasheed, S. A. Khan, A. Hassan and S. Safdar, "A Decision Support Framework for National Crop Production Planning," in IEEE Access, vol. 9, pp. 133402-133415, 2021, doi: 10.1109/ACCESS.2021.3115801.

[18] M. A. Z. Abidin, M. N. Mahyuddin and M. A. A. M. Zainuri, "Optimal Efficient Energy Production by PV Module Tilt-Orientation Prediction Without Compromising Crop-Light Demands in Agrivoltaic Systems," in IEEE Access, vol. 11, pp. 71557-71572, 2023, doi: 10.1109/ACCESS.2023.3293850.

[19] C. A. Martínez Félix, G. E. Vázquez Becerra, J. R. Millán Almaraz, F. Geremia-Nievinski, J. R. Gaxiola Camacho and Á. Melgarejo Morales, "In-Field Electronic Based System and Methodology for Precision Agriculture and Yield Prediction in Seasonal Maize Field," in IEEE Latin America Transactions, vol. 17, no. 10, pp. 1598-1606, October 2019, doi: 10.1109/TLA.2019.8986437.

[20] H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy and A. Elmahalawy, "Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," in IEEE Access, vol. 10, pp. 64671-64687, 2022, doi: 10.1109/ACCESS.2022.3181970.

[21] L. He, C. A. Coburn, Z. -J. Wang, W. Feng and T. -C. Guo, "Reduced Prediction Saturation and View Effects for Estimating the Leaf Area Index of Winter Wheat," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 3, pp. 1637-1652, March 2019, doi: 10.1109/TGRS.2018.2868138..

[22] Dr. Avinash Pawar. (2020). Development and Verification of Material Plasma Exposure Concepts. International Journal of New Practices in Management and Engineering, 9(03), 11 - 14. https://doi.org/10.17762/ijnpme.v9i03.90

[23] Dr. Nitin Sherje. (2020). Biodegradable Material Alternatives for Industrial Products and Goods Packaging System. International Journal of New Practices in Management and Engineering, 9(03), 15 - 18. https://doi.org/10.17762/ijnpme.v9i03.91