ISSN: 2769-545X

Research Article

Advances in Machine Learning & Artificial Intelligence

A Comparative Study of Ensemble Classification Algorithms for Crop Yield Forecasting

Normias Matsikira*, Gideon Mazambani and Martin Muduva

Chinhoyi University of Technology, Zimbabwe

***Corresponding Author**
Normias Matsikira, Chinhoyi University of Technology, Zimbabwe.

Submitted: 2024, Dec 30; **Accepted:** 2025, Feb 05; **Published:** 2025, Feb 20

Citation: Matsikira, N., Mazambani, G., Muduya, M. (2025). A Comparative Study of Ensemble Classification Algorithms for Crop Yield Forecasting. *Adv Mach Lear Art Inte, 6*(1), 01-05.

Abstract

This study explores the application of ensemble learning techniques to improve predictive model accuracy. It focuses on combining classifiers to outperform individual models using structured and unstructured data from agricultural datasets. Artificial neural networks (ANNs) and ensemble methods were used to increase deep neural network efficiency. Experiments with different network structures, training iterations, and topologies were conducted, evaluating measures like sensitivity and specificity. The research also includes predicting crop yields using ensemble classification algorithms, comparing accuracy with conventional methods. The study highlights the importance of crop yield prediction for agricultural management and discusses the benefits of ensemble methods. Results show that Random Forest, XGBoost, AdaBoost, and ANNs perform well in predicting crop yields as compared to the other algorithms. This research contributes to understanding the impact of weather patterns and genotype on crop yields.

Keywords: Ensemble Classification Algorithms, Crop Yield Forecasting, Agriculture, Machine Learning, Agricultural Productivity, Crop Production

1. Introduction

This research provides an integrated approach of applying innovative ensemble learning techniques that has the potential to increase the overall accuracy of predictive models. The main aim of generating combined classifier ensembles is to improve the prediction accuracy in comparison to using an individual classifier. A combined predicting ensemble can improve the prediction results by compensating for the individual algorithms weaknesses in certain areas and benefiting from better accuracy of the other ensembles in the same area. Actual structured and unstructured data sets from industry are utilized during the research process, analysis and subsequent model evaluations.

Artificial neural networks have been successfully applied to a variety of machine learning problems, including stock market prediction, image recognition, semantic segmentation, and machine translation. However, few studies like fully investigated ensembles of machine learning algorithms [1]. In this work, we investigated multiple widely used ensemble methods, including unweighted averaging, majority voting, the Bayes Optimal Classifier, and the (discrete) Super Learner, for the purpose of increasing their efficiency in doing tasks, with deep neural networks as candidate algorithms.

In this study, different algorithms have been proposed for designing predicting ensemble combiners. The different methods are studied in detail and analysed using different datasets and databases. The combiner methods are compared empirically with several stand-alone classifiers using neural network algorithms. Different types of neural network topologies are used to generate different models.

The study designs several experiments, with the candidate algorithms being the same network structure with different model checkpoints within a single training process, networks with same structure but trained multiple times stochastically, and networks with different structure. Standard accuracy measures will be used, namely accuracy, sensitivity, specificity and area under the curve, in addition to training error accuracies such as the mean square error.

In this investigation, scientists will evaluate how well several ensemble classification algorithms predict crop yields. Crop yields and weather patterns from the past will be used in the study. To determine the elements that are most important for forecasting future yields, the data will first be analysed. Then, using a subset of the data, a number of ensemble algorithms will be trained and evaluated. Based on fresh meteorological information, the most precise algorithm will be chosen and applied to forecast future crop yields. The accuracy of these forecasts will then be assessed and contrasted with more conventional single-algorithm methods. The overall goal of this work is to enhance our knowledge of the relationship between weather patterns and crop productivity and to discover the best ensemble algorithm for predicting crop yields.

2. Background to the Study

A need for quantitative Statistical crop yield forecast outlooks has been felt for quite some time. With more than 345 million people experiencing severe food insecurity in 2023, the present global hunger and malnutrition epidemic is incredibly large [2]. A beginning towards its realization has been made by undertaking a study of past crop yield in relation to meteorological parameters, principally rainfall and temperature. Based on weather studies, crop yield forecast models are prepared for estimating yield much before actual harvest of the crops. By use of empirical machine learning models using correlation and regression techniques crops yield are forecast on an operational basis for the country. Meteorological parameters at various crop growth stages along with technological trends are used in the models. For the production of food on a worldwide scale, crop yield prediction is crucial. To improve national food security, policymakers must make timely import and export choices based on reliable forecasts [3].

Crop yield forecasting is an important responsibility for farmers, government representatives, and other agricultural stakeholders since it aids in planning food supplies, lowering market risks, and controlling the consequences of climate change. However, the process of predicting crop yields is inherently complicated because it takes into account a number of different factors, including weather, agronomic practices, soil quality, and more. The challenge in predicting is further increased by the necessity for accuracy and the accessibility of dependable data. In light of the fact that machine learning techniques provide automated data-driven methods for handling complicated and dynamic data, they have recently been embraced for crop production forecasting. A well-liked machine learning method called ensemble techniques combines several models to increase the overall prediction accuracy. The aforementioned study compares the efficacy of multiple ensemble classification algorithms for crop yield predictions, including Random Forest, RNN and AdaBoost.

3. Literature Review

3.1. Overview of Crop Yield Forecasting

Crop yield forecasting has traditionally been a crucial component of agricultural planning and management [4]. Crop yield forecasting techniques have developed over time, including new scientific knowledge and cutting-edge technology. Crop yield predictions have historically been made using arbitrary techniques that rely on farmers' knowledge and intuition. This required keeping an eye on physical aspects including the crop's health, development stage, and local weather patterns. Then, farmers would forecast the anticipated harvest using their expertise and experience. Objective procedures were used in addition to the subjective ones. These included crop simulation models that employed a range of data sources, including soil, crop, and weather data, statistical regression models that used historical yield data, and agrometeorological models that used weather data to predict yields.

However, these age-old techniques had their drawbacks. They frequently made assumptions based on sparse data and relied on methods that might not always be accurate. Examples include the assumption made by agrometeorological models that there is always a direct correlation between weather factors and agricultural yield. Similar assumptions were made in statistical regression models, which may not hold true given changes in farming methods, technological advancements, and climate change. On the other hand, crop simulation models were frequently sophisticated and required a lot of data, which wasn't always available, especially in developing nations.

The investigation of machine learning approaches for crop yield forecasting resulted from these limitations. The employment of ensemble classification algorithms is one such method [5]. To provide a final prediction, these algorithms aggregate the results of various models. A number of weak learners can combine to become a strong learner, according to the theory behind ensemble methods [6]. Therefore, even if each model has certain flaws, when the models are integrated, these flaws can balance each other out and produce a prediction that is more accurate.

3.2. Ensemble Classification Algorithms

Liu (2014) defined ensemble classification algorithms as machine learning models that use predictions from many models to make predictions. The basic idea behind how they work is that fresh examples are classified using a combination of base classifiers created from training data. Argues that ensemble approaches work to improve prediction performance by utilizing the diversity among the basic classifiers, which is brought about by utilizing several learning algorithms or training data subsets [7].

According to, ensemble approaches have two main benefits: increased prediction accuracy and robustness [8]. Ensemble approaches, which combine many models, can take advantage of each model's advantages while making up for its shortcomings, resulting in a prediction that is more accurate in the aggregate. Additionally, because ensemble approaches average the prediction errors across various models, they are resistant to overfitting and noise [9]. Provide additional evidence that ensemble models are better suited to handle big, complicated datasets because they divide the data into smaller, more manageable portions [1].

3.3. Using Ensemble Algorithms to Predict Crop Yield

Due to their great predictive performance, ensemble classification

algorithms have been extensively used in the area of crop production forecasting. The Random Forest algorithm, developed by (Breiman, 1996b), stands out among the well-known ensemble algorithms for its efficiency and simplicity [10]. It works by building numerous decision trees and producing a class that is the mode of the classes of each tree individually.

Another effective ensemble technique that is frequently applied in agricultural production prediction is gradient boosting. According to (Friedman et al., 1999), new models are fitted to provide a better approximation to the gradient of the loss function [11]. This process constructs an additive model in a forward stage-wise way. Bagging is an ensemble method that modifies the training data by uniformly sampling and using replacement [10,12]. It lowers a base learning algorithm's variance, frequently resulting in a significant improvement in stability and accuracy [12].

3.4. Data

Participants in the 2018 Syngenta Crop Challenge were tasked with predicting the performance of corn hybrids in 2017 in various regions in India using real-world data [13]. The dataset comprised 2,267 experimental hybrids that were spread across India between 1997 and 2014 in 2,247 different places. This dataset was selected because it has been anonymised (de-identified) and is accurate, confidentiality has been guaranteed [13]. The dataset had too many attributes, we had to exclude some of the attributes before analysis because they had insignificant impact on the crop yield. In this study, we used the guided backpropagation method, which backpropagates the positive gradients to find input variables that optimize the activity of our interesting functions [14].

All of the places were spread out across India. This was one of the biggest and most complete datasets for study in crop yield statistics that was publicly available, allowing for the deployment and validation of the suggested ensemble machine learning algorithms. The yield performance dataset included 148,452 samples for various hybrids planted in various years and locations, together with the observed yield, check yield (average yield across all hybrids at the same location), and yield differential. Yield difference, which measures how well a hybrid performs in comparison to other hybrids at the same area, is the difference between yield and check yield [15]. Participants in the Syngenta Crop Challenge received normalized and anonymised meteorological data. With no training data overlap, each validation sample included a unique hybrid and location combination.

3.5. Related Studies

Chen (2017) proposed an entropy-based combination prediction model for predicting unit crop yield. The researchers combined the grey forecasting model and radial basis function neural network forecasting models to improve the accuracy of predictions. This combined model is considered less risky, practical, intuitive, and feasible.

Dehzangi et al (2017) developed an ensemble method using different classifiers such as Adaboost.M1, Logitboost, Naive Bayes, Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM) for predicting protein structural class. They used auto-correlation based feature extraction techniques to achieve better results.

Kumar et al (2020) presented a novel machine learning model for solving the crop selection problem. They introduced a method called Crop Selection Method (CSM) to identify the crop selection for a specific region. Their conclusion was that proper crop selection using CSM increases the net crop yield.

Rahman et al (2020) developed a machine learning model for predicting rice production in Bangladesh, where the soil condition is not homogeneous. They trained their models using the correlation between previous environmental climate and crop yield rate. Finally, they compared the performance of different models to assess their accuracy.

Deepti (2020) developed rainfall forecasting models based on algorithms such as Classification and Regression Tree, Naive Bayes, K Nearest Neighbors, and 5-10-1 Pattern Recognition Neural Network. The corresponding techniques yielded accuracy results of 80.3%, 78.9%, 80.7%, and 82.1%, respectively.

Meshram et al (2018) focused on the challenges in weather prediction. They used meteorological data obtained from the Indian Meteorological Department (IMD) for their analysis and prediction of rainfall. Bayesian data mining techniques were utilized, and the results showed good prediction accuracy with moderate computing resources using the Bayesian approach. These studies highlight various approaches and techniques used in forecasting and prediction tasks related to crop yield, protein structural class, rainfall, and weather.

5. Methodology

Sklearn, pandas, numpy, and matplotlib were used to explore the data and implement the study's algorithms. The flow of the algorithm implementation is as shown in Figure 1.

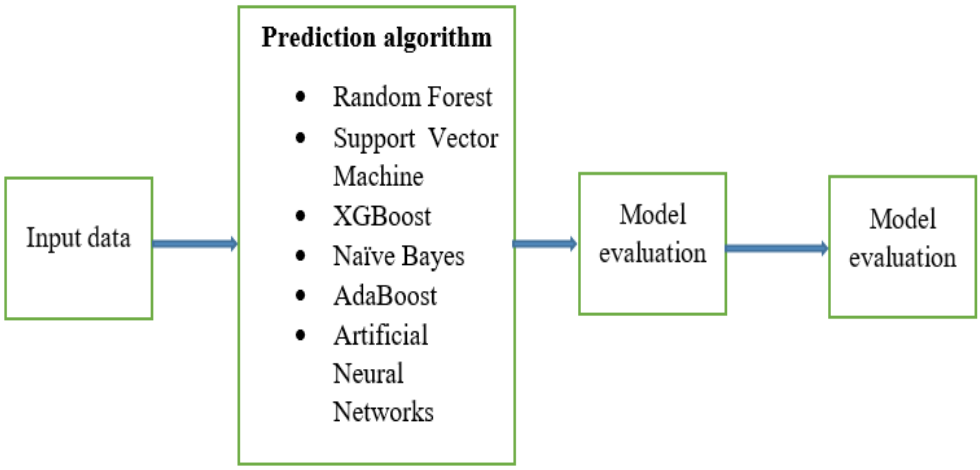

Figure 1: Methodology Flow

6. Results

Table 1 presents the consolidated results of the metrics of the algorithms used while Figure 2 shows a visualisation of the same metrics.

MODEL	ACCURACY	R2-SCORE	Mean Absolute Error
Random Forest	0.9576	0.9589	0.8953
Support Vector Machine	0.6512	0.5560	4.5219
XgBoost	0.8864	0.9654	3.3521
Naïve Bayes	0.7101	0.5030	0.8950
Adaboost	0.9434	0.9572	0.2722
Artificial Neural Networks	0.9688	0.9310	0.2722

Table 1: Model Results

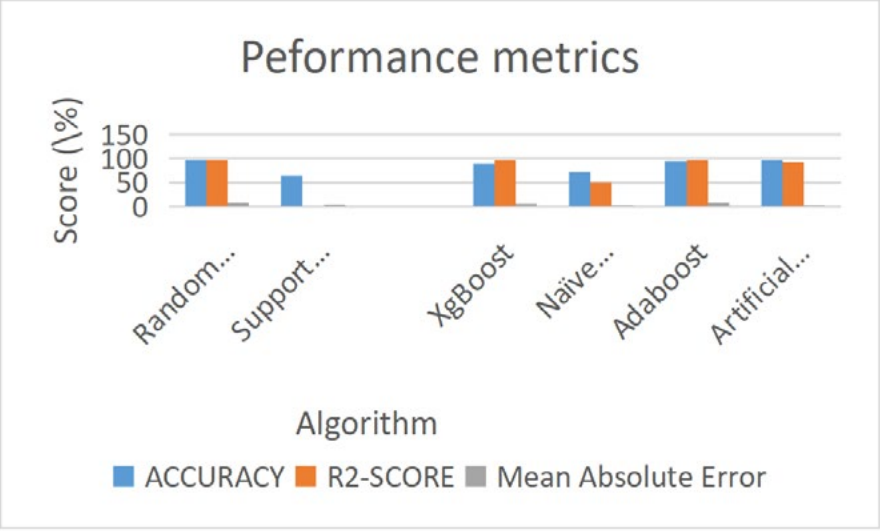

Figure 2: Model Results Visualisation

The results suggests that Random Forest, XGBoost, AdaBoost, and Artificial Neural Networks demonstrate good performance in predicting crop yield, with high accuracy values ranging from 0.9434 to 0.9688. This indicates that these models can effectively classify crop yields with a high level of accuracy, which is crucial for accurate yield prediction.

The R2-score results show that the XGBoost achieves the highest score of 0.9654, implying that it can explain a significant proportion of the variance in crop yield. Random Forest and AdaBoost also perform well, with R2-scores of 0.9589 and 0.9572, respectively. These models can capture patterns and relationships in the data to a great extent.

Mean Absolute Error (MAE) measures the average magnitude of prediction errors. Artificial Neural Networks recorded the lowest MAE at 0.2722, indicating that it generally makes more precise predictions of crop yield compared to the other models. AdaBoost and Random Forest have slightly higher MAE values of 0.8950 and 0.8953, respectively, while XGBoost has an MAE of 0.6670. These values represent the average deviation of the predictions from the actual crop yield values.

Support Vector Machine and Naïve Bayes models exhibit relatively lower accuracy and R2-scores. Support Vector Machine has an accuracy of 0.6512, while Naïve Bayes has an accuracy of 0.7101. However, the R2-scores for these models are not provided, making it challenging to fully evaluate their performance. Additionally, Naïve Bayes has a higher mean absolute error of 3.3521, suggesting less accurate predictions compared to other models. Based on these findings, the models that perform well in predicting crop yield are Random Forest, XGBoost, AdaBoost, and Artificial Neural Networks. Considering their high accuracy, R2-scores, and relatively low mean absolute error, it is recommended to utilize these models for crop yield prediction tasks.

7. Conclusion

Using the Indian agricultural dataset, the study employed an ensemble machine learning approach for crop yield prediction that outperformed the other algorithms utilizing massive amounts of data. Based on genotype and environmental information, the strategy used ensemble methods to predict crop yields (including yield, check yield, and yield differential). From historical data, the carefully crafted ensemble methods were able to identify nonlinear and complex relationships between genes, environmental factors, and their interactions, and they were able to reasonably predict the yields of new hybrids planted in unfamiliar places with known weather conditions. The model's performance was discovered to be rather sensitive to the accuracy of the weather forecast, which indicated the significance of weather prediction methods. One major limitation of ensemble methods used in this study is that it is extremely vulnerable to noise data and outliers, it is also more prone to overfitting that other algorithms.

Funding Declaration

No funding was provided to all authors during the course of this research.

All authors are based in Harare, Zimbabwe.

Ethics and consent to participate declarations: not applicable.

References

1. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research, 11*, 169-198.
2. Laganda, G. (2023). Responding to loss and damage in food systems. *Nature Food, 4*(2), 133-134.
3. Horie, T., Nakagawa, H., Centeno, H. G. S., & Kropff, M. J. (1995). The rice crop simulation model 4 SIMRIW and its testing. *Modeling the impact of climate change on rice production in Asia. Wageningen University*, 95-139.
4. Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 125-130). IEEE.
5. Medar, R., Rajpurohit, V. S., & Shweta, S. (2019, March). Crop yield prediction using machine learning techniques. In *2019 IEEE 5th international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE.
6. Adeva, J. J. G., Beresi, U. C., & Calvo, R. A. (2005). Accuracy and diversity in ensembles of text categorisers. *CLEI Electronic Journal, 8*(2), 1-1.
7. Zhou, S., Chen, Q., & Wang, X. (2010, August). Active deep networks for semi-supervised sentiment classification. In *Coling 2010: Posters* (pp. 1515-1523).
8. Kuncheva, L. I. (2004). Classifier ensembles for changing environments. *In Multiple Classifier Systems: 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004. Proceedings 5* (pp. 1-15). Springer Berlin Heidelberg.
9. Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine, 6*(3), 21-45.
10. Breiman, L. (1996). Bagging predictors. *Machine learning, 24,* 123-140.
11. Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999, August). Learning probabilistic relational models. *In IJCA*I (Vol. 99, pp. 1300-1309).
12. Breiman, L. (1996). Bagging predictors Machine Learning 24 (2), 123-140 (1996) 10.1023. *A: 1018054314350.*
13. Johnson, M. A., Spore, T. J., Montgomery, S. P., Weibert, C. S., Garzon, J. S., Hollenbeck, W. R., ... & Blasi, D. (2018). Syngenta Enhanced Feed Corn (Enogen) containing an alpha amylase expression trait improves feed efficiency in growing calf diets. *Kansas Agricultural Experiment Station Research Reports, 4*(1), 16.
14. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806.*
15. Stephens, A. D., Banigan, E. J., Adam, S. A., Goldman, R. D., & Marko, J. F. (2017). Chromatin and lamin A determine two different mechanical response regimes of the cell nucleus. *Molecular biology of the cell, 28*(14), 1984-1996.