

Research

Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture

O. Sri Nagesh¹ · Raja Rao Budaraju² · Shriram S. Kulkarni³ · M. Vinay⁴ · Samuel-Soma M. Ajibade⁵ · Meenu Chopra⁶ · Malik Jawarneh⁷ · Karthikeyan Kaliyaperumal⁸

Received: 3 November 2023 / Accepted: 10 April 2024

Published online: 25 April 2024

© The Author(s) 2024 [OPEN](#)

Abstract

Due to the limited availability of natural resources, it is essential that agricultural productivity keep pace with population growth. Despite unfavorable weather circumstances, this project's major objective is to boost production. As a consequence of technological advancements in agriculture, precision farming as a way for enhancing crop yields is gaining appeal and becoming more prevalent. When it comes to predicting future data, machine learning employs a number of methods, including the creation of models and the acquisition of prediction rules based on past data. In this manuscript, we examine various techniques to machine learning, as well as an automated agricultural yield projection model based on selecting the most relevant features. For the purpose of selecting features, the Grey Level Co-occurrence Matrix method is utilised. For classification, we make use of the AdaBoost Decision Tree, Artificial Neural Network (ANN), and K-Nearest Neighbour (KNN) algorithms. The data set that was used in this study is simply a compilation of information about a variety of topics, including yield, pesticide use, rainfall, and average temperature. This data collection consists of 33 characteristics or qualities in total. The crops soya beans, maize, potato, rice, paddy, wheat, and sorghum are included in this data collection. This data collection was made possible through the collaboration of the Food and Agriculture Organisation (FAO) and the World Data Bank, both of which make their data available to the public. The AdaBoost decision tree has achieved the highest level of accuracy possible when used to anticipate agricultural yield. Both the accuracy rate and the recall rate are quite high at 99 percent.

Article highlights

- This article presents boosting and optimum feature selection based efficient machine learning technique for accurate prediction of crop yield
- Proposed methodology uses Gray Level Co-occurrence Matrix technique is used to select features. The AdaBoost Decision Tree, ANN, and KNN algorithms are used for classification.
- The performance of AdaBoost machine learning is more accurate for crop yield prediction. Adaboost has achieved 98 percent accuracy, 99 percent Recall rate.

✉ Karthikeyan Kaliyaperumal, karthikeyan@ambou.edu.et; O. Sri Nagesh, nagesh.osri@gmail.com; Raja Rao Budaraju, rajaraob@yahoo.com; Shriram S. Kulkarni, kulk.shri@gmail.com; M. Vinay, vinay.m@christuniversity.in; Samuel-Soma M. Ajibade, asamuel@ticaret.edu.tr; Meenu Chopra, drmeenuchopravips@gmail.com; Malik Jawarneh, ieeemtech@gmail.com | ¹Department of CSE, Anurag University, Hyderabad, India. ²Oracle Inc., Austin, USA. ³Department of Information Technology, Sinhgad Academy of Engineering, SavitribaiPhule Pune University, Pune, India. ⁴Computer Science, CHRIST (Deemed to be University), Bengaluru, India. ⁵Department of Computer Engineering, Istanbul Ticaret University, Istanbul, Turkey. ⁶Vivekananda Institute of Professional Studies, New Delhi, India. ⁷Faculty of Computing Sciences, Gulf College, Muscat, Oman. ⁸IoT - HH Campus, AMBO University, Ambo, Ethiopia.



Keywords *Boosting, Crop Yield Prediction · Feature Selection · Gray Level Co-occurrence Matrix · AdaBoost Decision Tree · Accuracy*

1 Introduction

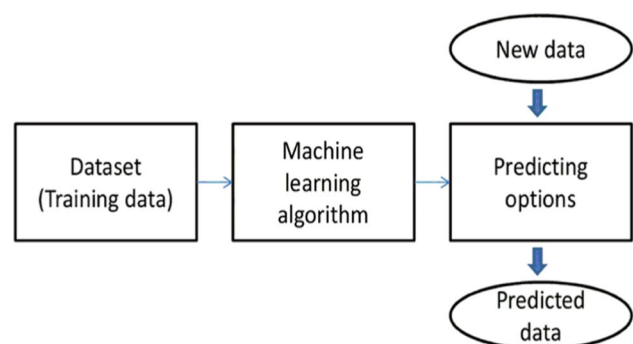
Because there is a limited supply of natural resources, it is imperative that agricultural production maintain up with rising population levels. In spite of the unfavorable weather conditions, the primary goal of this project is to increase productivity. As a result of developments in agricultural technology, the use of precision farming as a method for increasing crop yields is gaining popularity and is becoming more widespread. When it comes to forecasting future data, machine learning involves a variety of approaches, including the construction of models and the learning of prediction rules based on historical information (ML). The capability of a computer to learn from experience is referred to as “machine learning.” When these two methodologies are combined, the development of intelligent robots that are fully automated and equipped with artificial intelligence is feasible. We may argue that big data and data mining are mutually supportive of one another given that they are complementary to one another [1, 2]. Machine learning model for prediction work is shown in Fig. 1.

The fact that the vast majority of datasets that are compiled from a variety of sources include errors is a significant limitation that applies to machine learning algorithms. Many difficulties may arise for machine learning algorithms when they must work with data that is incomplete, duplicated, has outliers, or is otherwise inconsistent.

In the field of computer science, the term for self-teaching systems is “machine learning.” This is because these systems are able to learn from a data set without being specifically built to do so, which is the reason behind this result. A group of steps that, when carried out on computers, make it possible to make precise predictions on what will happen in the future. The objective of machine learning is to have a computer programme learn from new data as it is entered into the system so that it may construct an algorithm that can anticipate the output based on statistical calculations. This is accomplished through the use of the term “machine learning.” As a direct result of this, data samples are analysed by means of machine learning in order to find patterns and establish decision rules for the goal of developing a predictive model that can be used to anticipate future data. This can be done so that the model can be used to anticipate the data that will be coming in the future. The more experience they get, the more likely it is that predictive models will eventually be able to train themselves and come to their own judgements about specific conditions without needing any aid from a person in the process. There is historical data accessible that relates to crop yields. The historical crop yield data can be used as a training set for machine learning techniques, which can then be used to forecast future crop yield data. Both supervised and unsupervised forms of machine learning are possible. These are the two main categories. A significant number of algorithms may be found in each of these fields [3].

This article explains the techniques of machine learning as well as an automatic crop yield forecast model that is based on optimised feature selection. The Grey Level Co-occurrence Matrix technique is utilised in the process of feature extraction. The AdaBoost Decision Tree, the ANN Algorithm, and the KNN Algorithm are used in the classification process. The data set that was employed in this investigation is essentially a collection of information pertaining to precipitation, pesticides, yield, and the typical temperature. This data set includes a total of 33 characteristics or qualities. This data set includes information pertaining to crops such as soybeans, maize, potato, rice, paddy, wheat, and sorghum. This data collection was compiled by collecting information from databases maintained by the Food and Agriculture Organisation

Fig. 1 Machine Learning for Prediction Task



(FAO) and the World Data Bank (WDB), both of which make their databases accessible to the public. The AdaBoost algorithm now achieves the maximum accuracy possible when predicting crop yield.

Section 2 contains literature review of existing techniques for data preprocessing and crop yield prediction. Section 3 presents boosting machine learning based framework for crop yield prediction. Section 4 presents result analysis. Section 5 depicts conclusion and future work.

2 Literature survey

ID3 approach makes use of both entropy and information gain measurements. At each stage of the analysis that is performed on the initial dataset, the entropy of the functional features is computed as part of the analysis. When dividing up datasets into subgroups, the characteristics that have the most information gain and the lowest entropy are the ones that are chosen (split attribute). Repeat the process with each new set of data until there is a perfect fit for all of the information in the category that was established beforehand. By utilising the branch in question's final subset, it is possible to differentiate between the terminal and non-terminal nodes of a branch. In this case, the split property is used to define the non-terminal node, while the class labels are used to describe the terminal node. With the assistance of a decision tree model that is constructed on the ID-3 data set, it is possible to identify and forecast heart disease with a greater degree of accuracy. When creating prediction models for large health datasets, decision tree modelling techniques such as Classification and Regression tree (CART) and ID-3 are widely utilised. Examples include. A tenfold cross-validation method is utilised in order to ensure the accuracy of the findings. Based on the data, it appears that decision tree classification strategies might be used to construct a prediction model that is both accurate and cost-effective. You might be able to analyse smaller datasets more quickly with the ID-3 approach, and the results could be more accurate. When it comes to things like electronic health records, for example, the constant modifications require a sizeable amount of additional storage space for the data [3, 4].

The author of [5] outlines the decision tree strategy to categorising data. Model overfitting is becoming increasingly prevalent in ID-3-based decision tree classification algorithms, which affects the reliability of categorising health data. Develop a schema for the classification of data in order to ensure that it may be appropriately organised. We use a top-down strategy, so we begin with the most important node, which is called the root, and work our way out to the ones at the very end (the leafs). Entropy measurements are used both to choose a meaningful feature and to partition the input data into subsets in order to accomplish these two tasks. When making a choice, the factor that is given the greatest consideration is the property that has the largest normalised information gain. Its superior performance can be attributed to the elimination of unnecessary features from the branches. Because of this, it is effective at processing numeric data, including data that has values that are missing. When it comes to dealing with quantitative characteristics, the decision tree is put through a more rigorous examination.

According on the findings presented in [6], the decision tree C-5 was developed to address a number of the shortcomings of the C4.5 methodology. Calculations are able to be completed more quickly as a result of its ability to shrink the size of the tree while simultaneously using less storage space. Boosting is allowed, and any characteristics that are eliminated immediately from consideration for the classification are ones that are not vital to it. The training dataset is limited in order to ensure that the search will be focused on the most relevant information. A second battery of examinations is used in order to validate the final outcomes of the categorization. The severity of heart disease and its risk factors may be linked together via the use of association rules. In the C-5 decision tree approach, there are fewer rules to follow, but the accuracy of the results has also increased.

According to the author, one approach that is often used is decision trees. These trees are constructed using classification and regression trees, as detailed in [7]. Every single base nodule is equipped with its own inputs and bases. The value that is provided for the input attribute has to be in the form of a number. It is possible to create predictions based on the output variable that is represented by the leaf node. The dataset is classified more accurately with the help of a statistical model that was created using discriminant analysis. Both in terms of qualities that are categorical and those that are continuous, it is a great fit.

The authors of [8] claim that they are able to categorise and forecast data via the use of ensemble learning. During the training phase, decision trees are constructed, and regression methods are used to make predictions for each tree's potential outcomes. Because of the model's reduced standard deviation and the simplicity with which it links the various features of the input data, it is possible to make predictions with a lower margin of error. As a result of how difficult they are to understand at first, there was a distinct lack of enthusiasm about random forest classification methods.

The statistical classification strategy that is based on Bayes's theorem is referred to as Bayesian classification, as stated by the authors [9]. It is presumed that a probabilistic model will be used, which performs well with numerical and categorical datasets alike. The other methods of categorisation take far more time than this one does. The categorization model is constructed by applying a predetermined set of methods and criteria to each step of the process. Assumes that each and every one of the capabilities of the system are included into the product. In this setting, the regular use of useful classification algorithms that are able to handle even the most complicated and big datasets may be found.

According to the findings of a research [10], the most accurate method for identifying patterns and classifying data is to make use of the K-nearest neighbour technique. Distance functions or measurements of similarity are the foundation for the k-nearest neighbour algorithms. The storage of an instance's current state may be used to categorise instances according to the degree of similarity between them. When using instance-based learning, classification may be completed in a shorter amount of time. A recently created instance of the dataset has been categorised by the overwhelming majority of its immediate neighbours. The first thing that has to be done is to establish the value of k by using data from both the training set and the testing set, and then figure out how far away the two sets are from one another.

In order to construct a reliable model for accurately forecasting the results of the input datasets, predictive analytic approaches are used. Using the appropriate data mining tools, rules, and correlations, it is possible to construct predictive models. It increases the usefulness of the data and generates predictions about the way in which an event will unfold in the future. Classification is a term used in predictive analysis, and it relates to the process of appropriately mapping and categorising the information that is presented. The use of data to create projections about the course of future events is the focus of the kind of data analysis known as predictive analysis. The generation of predictions is accomplished by the use of statistical and machine learning strategies, in addition to data mining and artificial intelligence. It analyses data in its raw or unprocessed form and searches for patterns in human behaviour in order to extract knowledge that may be put to practical use from the data. This metric is used to determine how likely it is that each individual component will fulfil its forecast. The application of this technique enables one to generate predictions about the future that are based on solid evidence. In order to forecast future occurrences, we make use of predicted values and dependant variables. The ability of a system to make better decisions is one of the areas that may be enhanced using prescriptive analytic methods.

Research suggests that machine learning will play an increasingly significant role in predictive analysis in the future. For the purpose of making predictions, predictive analysis employs a five-step method. To begin, the scope of the project has to be defined. After the scope of the project and its goals have been defined, the next step is to identify the data sets that will be needed to make forecasts. In the second stage, data is gathered from a variety of different sources. Sanitizing the data involves removing any noise or unnecessary information from it before it can be used to extract meaningful information. In the third step, assumptions and hypotheses are put through their paces by employing statistical models. A predictive model is constructed when it has been established that the assumptions are correct. The deployment procedure is the very final step that has to be completed. The model may provide real-time forecasts, which can subsequently be used in order to make more informed judgments. Model monitoring is the fifth phase of predictive analysis. Throughout the whole of this phase, the performance of the model is subjected to on-going assessment. If you want your predictive analysis to be more accurate and completed more quickly, you should follow these seven stages [11].

3 Methodology

In the following section, we will discuss various machine learning techniques and an autonomous crop yield forecast model that is based on optimised feature selection. The Grey Level Co-occurrence Matrix technique is utilised in the process of feature extraction. The GLCM method is utilised because it is able to recover the characteristics of the texture without causing any disruption to the interpixel connections that exist between the various individual pixels. The AdaBoost Decision Tree, the ANN Algorithm, and the KNN Algorithm are used in the classification process. AdaBoost excels when it comes to enhancing the efficiency with which decision trees deal with binary classification issues. In principle, it ought to be feasible to make use of it to improve the overall performance of any approach of machine learning. The data set that was employed in this investigation is essentially a collection of information pertaining to precipitation, pesticides, yield, and the typical temperature. This data set includes a total of 33 characteristics or qualities. This data set includes information pertaining to crops such as soybeans, maize, potato, rice, paddy, wheat, and sorghum. This data collection was compiled by collecting information from databases maintained by the Food and Agriculture Organisation (FAO) and the World Data Bank (WDB), both of which make their databases accessible to the public. The AdaBoost algorithm now

achieves the maximum accuracy possible when predicting crop yield. Figure 2 illustrates proposed GLCM and machine learning based framework for accurate crop yield prediction.

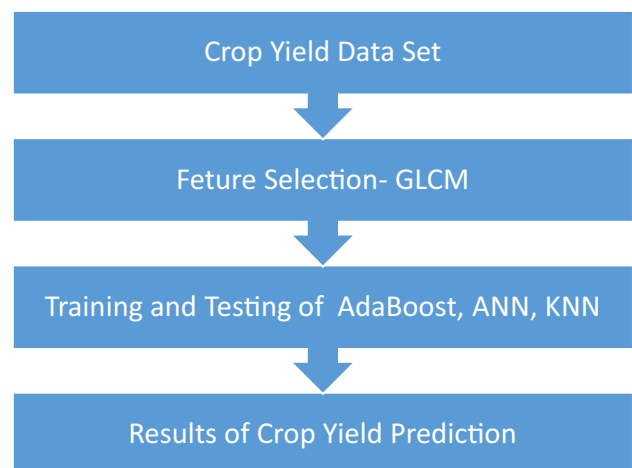
The field of image processing makes use of a technology known as feature extraction, which may be implemented to cut down on the quantity of data. In order to achieve this goal, unnecessary or redundant dimensions from a collection of feature subsets must first be removed. The GLCM approach is used because it can restore the properties of the texture without disrupting the interpixel connections between the individual pixels [12].

The K-nearest neighbour method [13] is the one that provides the most reliable results when it comes to the recognition of patterns and the classification of data. Distance functions and similarity metrics are the foundation upon which k-nearest neighbour algorithms are constructed. With the assistance of the similarity measure, it is simpler to categorise newly occurring instances and to monitor the present standing of all formerly occurring ones simultaneously. The utilisation of instance-based learning makes effortless categorization possible. The majority of votes from the classes that are adjacent to the one being categorised are used for classifying the new instance of the dataset. The data from the training sets and the testing sets are used in the calculation of the distance measure. In order to get things rolling, the first thing that needs to be done is to calculate the value of k and determine the distance that separates one instance from the next.

Many of the functions of a Neural Network are analogous to those of the human brain. The core processing units of a natural nervous system, known as neurons, serve as a source of inspiration for the corresponding components of an artificial neural network, also known as an ANN. These are the fundamental components that go into making up the signalling infrastructure of the nervous system. Every single neuron is constructed from its very own specialised cell, each of which is capable of its own unique set of tasks. Components of theoretical circuits that are able to carry out computations have been modelled after the architectural structures of biological networks. The operation of a single neuron must first be comprehended in order to make progress towards comprehending the functioning of the brain as a whole. An artificial neural network, often known as an ANN, is a method of data processing that makes use of a connectionist approach. This system is built on top of an artificial neural network as its foundation. There are certain points of comparison that can be made between this idea and the axons and dendrites that are found in the neurological system. The capability of the model to self-organize and change is what truly sets it apart from other models, despite its adaptability, scalability, and capacity to collaborate computationally with other models. An artificial functional model of a biological neuron is made up of the three components listed below: In order to simulate the connections that can be found in genuine neurons, weights are utilised. Both natural and artificial connections can exist between neurons in the brain. The form of synapse that is found in nature is by far the most common. Every one of the synaptic connections between artificial neurons is "weighted," which means that each connection is given a certain numerical value. When two variables are connected by weights, an inhibitory relationship is shown by a value with a negative value, and an excitatory relationship is indicated by a value with a positive value. Following the addition of all of the inputs comes the weighting process. In the jargon of computer science, this kind of combination is known as a linear combination. Utilising an activation function allows for the output amplitude to be subjected to a variety of modifications. For example, if the output numbers are in the range of 0 to 1 or -1 to 1, any value inside either of those bands is considered to be valid [14].

When Freund and Schapire first developed their technique, they gave it the moniker "AdaBoost M1," which was a moniker that they had come up with themselves. It is more usual to hear the term "discrete AdaBoost" these days because

Fig. 2 Boosting and optimum feature selection based efficient machine learning technique for accurate prediction of crop yield



it is typically used for classification rather than regression. This is one of the reasons why the name is more frequent. It should come as no surprise that the name has some bearing on this. We use a method known as Adaboost M1 class boost in order to improve the overall performance of a classifier that deals with nominal data. The scope of the performance improvements is restricted to problems with the nominal class levels. When it comes to increasing the effectiveness of decision trees for dealing with binary classification problems, AdaBoost shines. In theory, it should be possible to utilise it to enhance the performance of any machine learning method. These models perform remarkably well in the classification job when compared with the possibility of chance. Consequently, the combination of AdaBoost with single-level decision trees is the strategy that has shown to be the most successful and widespread. These trees are frequently referred to as decision stumps [15], on account of the single classification that can be assigned to them and their little stature.

4 Result analysis

Data set [16] used in this study is basically a collection of details related to rainfall, pesticides, yield and average temperature. This data set consists of 33 attributes or features. This data set contains related to crops like- Soybeans, Maze, Potato, Rice, paddy, Wheat and Sorghum. This data set is built by collecting data from publicly available data from Food and Agriculture organization (FAO) and World data bank. Features are extracted using the Gray Level Co-occurrence Matrix algorithm. Classification is performed using the AdaBoost decision tree, ANN and KNN algorithms.

Five parameters accuracy, specificity, precision, F1 and recall are used in experimental analysis. The accuracy, precision, recall and specificity of AdaBoost decision tree, ANN and KNN algorithms is shown in Figs. 3, 4, Tables 1 and 2.

The Grey Level Co-occurrence Matrix technique is utilised in the process of feature extraction. The AdaBoost Decision Tree, the ANN Algorithm, and the KNN Algorithm are used in the classification process. The data set that was employed in this investigation is essentially a collection of information pertaining to precipitation, pesticides, yield, and the typical temperature. This data set includes a total of 33 characteristics or qualities. This data set includes information pertaining to crops such as soybeans, maze, potato, rice, paddy, wheat, and sorghum. This data collection was compiled by collecting information from databases maintained by the Food and Agriculture Organisation (FAO) and the World Data Bank (WDB), both of which make their databases accessible to the public. The AdaBoost algorithm now achieves the maximum accuracy possible when predicting crop yield. Accuracy of AdaBoost GLCM is 98 percent, Specificity is 97 percent, precision is 99 percent and recall rate is 98 percent.

From results, it is clear that the accuracy, sensitivity, precision, recall and specificity of AdaBoost decision tree is better than KNN and ANN algorithm. It is also observed that the accuracy of machine learning algorithms has increased by 5 percent approximately when features are selected by Gray Level Cooccurrence Matrix algorithm.

Fig. 3 Accuracy, Specificity, Precision, F1 and Recall of AdaBoost Decision Tree, ANN, KNN with GLCM Feature Selection for Crop Yield Prediction

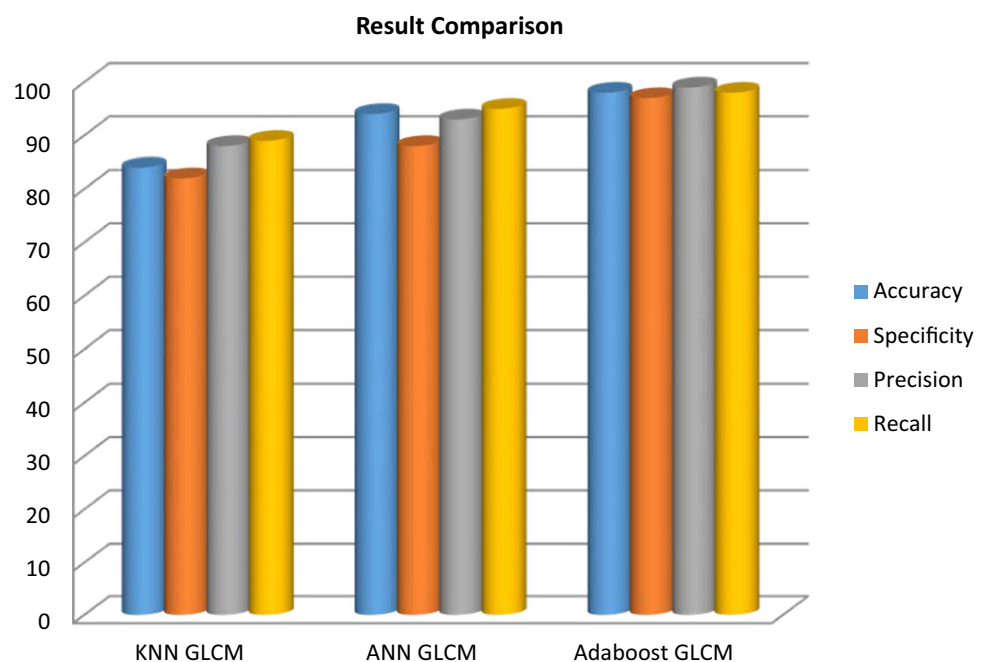


Fig. 4 Accuracy, Specificity, Precision, F1 and Recall of AdaBoost Decision Tree, ANN, KNN without GLCM Feature Selection for Crop Yield Prediction

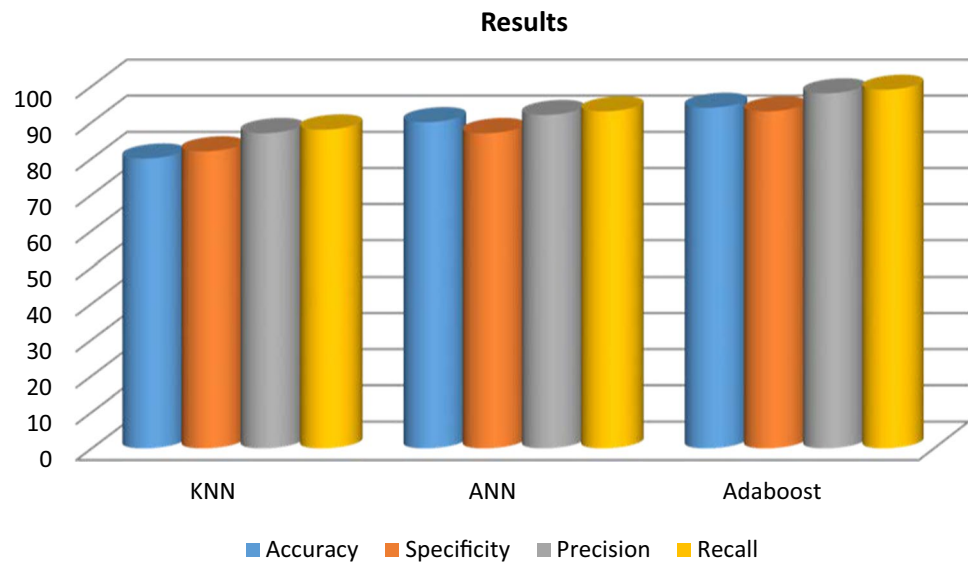


Table 1 Result comparison of classifiers with GLCM

Parameters	AdaBoost GLCM	ANN GLCM	KNN GLCM
Accuracy	98	94	84
Specificity	97	88	82
Precision	99	93	88
Recall	98	95	89
F1 Score	99	95	86

Table 2 Result comparison of classifiers without GLCM

Parameters	AdaBoost	ANN	KNN
Accuracy	94	90	80
Specificity	93	87	82
Precision	98	92	87
Recall	99	93	88
F1	95	92	84

5 Conclusion

Because of the finite amount of natural resources that are available, it is imperative that agricultural output maintain up with the expansion of the human population. Increasing output is the primary goal of this initiative, and it will be accomplished in spite of unfavourable weather conditions. As a result of technical advances in agriculture, precision farming as a method for increasing crop yields is gaining popularity and is becoming more commonplace. This is a direct consequence of these innovations. When it comes to forecasting future data, machine learning utilises a variety of different strategies, some of which include the building of models as well as the acquisition of prediction rules based on previous data. In this study, we examine several techniques to machine learning, as well as an automated agricultural yield projection model based on selecting the most relevant features. For the purpose of selecting characteristics, the Gray Level Co-occurrence Matrix method is used. The algorithms CNN, ANN, and KNN are used for carrying out classification. The data collection that was used in this study is simply a compilation of information about a variety of topics, including yield, pesticide use, rainfall, and average temperature. This data collection consists of 33 characteristics or qualities in total. The crops soya beans, maize, potato, rice, paddy, wheat, and sorghum

are included in this data collection. This data collection was made possible via the collaboration of the Food and Agriculture Organization (FAO) and the World Data Bank, both of which make their data available to the public. The agricultural output predictions made by CNN are now among the most accurate in the industry.

Author contributions Dr. O. Sri Nagesh and Dr. Karthikeyan Kaliyaperumal designed the model and the computational framework and analysed the data. Dr. Shriram S. Kulkarni and Vinay M carried out the implementation. Samuel-Soma M. Ajibade performed the calculations. Dr. Meenu Chopra and Malik Jawarneh wrote the manuscript with input from all authors. Mr. Raja Rao Budaraju conceived the study and were in charge of overall direction and planning

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The data that support the findings of this study are openly available at [<https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset>], reference number [16].

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Bhanu KN, Jasmine HJ, Mahadevaswamy HS. Machine learning Implementation in IoT based intelligent system for agriculture. *Int Conf Emerg Technol (INCET)*. 2020;2020:1–5. <https://doi.org/10.1109/INCET49848.2020.9153978>.
2. Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: a comprehensive review. *IEEE Access*. 2021;9:4843–73. <https://doi.org/10.1109/ACCESS.2020.3048415>.
3. Raghuvanshi A, Singh U, Sajja G, Pallathadka H, Asenso E, Kamal M, et al. Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming. *J Food Qual*. 2022;2022:1–8. <https://doi.org/10.1155/2022/3955514>.
4. Abu SarwarZamani L, Anand KPR, Prabhu P, MateenButtar A, HarikumarPallathadka AR, NokobiDugbakie B. Performance of machine learning and image processing in plant leaf disease detection. *J Food Qual*. 2022;2022:7. <https://doi.org/10.1155/2022/1598796>.
5. M. Santhosh, M. D. Sai and S. Mirza. Ensemble deep learning model for wind speed prediction. 2020 21st National Power Systems Conference (NPSC), 2020, pp. 1–5, doi: <https://doi.org/10.1109/NPSC49263.2020.9331836>.
6. M. Utsumi, I. Shigemori and T. Watanabe. Forecasting Electricity Demand with Dynamic Characteristics Based on Signal Analysis and Machine Learning. 2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 2020, pp. 1049–1054, doi: <https://doi.org/10.23919/SICE48898.2020.9240281>.
7. S. Nandhini and J. Marseline K.S. Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–4, doi: <https://doi.org/10.1109/ic-ETITE47903.2020.312>.
8. C. M. Suneera and J. Prakash. Performance Analysis of Machine Learning and Deep Learning Models for Text Classification. 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1–6, doi: <https://doi.org/10.1109/INDICON49873.2020.9342208>.
9. Yang C, Prasher SO, Landry J, Ramaswamy HS, Ditommaso A. Application of artificial neural networks in image recognition and classification of crop and weeds. *Can Agric Eng*. 2000;42(September):147–52.
10. Kaul M, Hill RL, Walthall C. Artificial neural networks for corn and soybean yield prediction. *Agric Syst*. 2005;85(1):1–18.
11. Leemans V, Magein H, Destain M-F. On-line fruit grading according to their external quality using machine vision. *Biosyst Eng*. 2002;83(4):397–404.
12. M. Benco, P. Kamencay, M. Radilova, R. Hudec and M. Sinko. The Comparison of Color Texture Features Extraction based on 1D GLCM with Deep Learning Methods. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, pp. 285–289, doi: <https://doi.org/10.1109/IWSSIP48289.2020.9145263>.
13. Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. *Int Conf Intell Comput Control Syst (ICCS)*. 2019;2019:1255–60. <https://doi.org/10.1109/ICCS45141.2019.9065747>.
14. N. Darapaneni et al. Handwritten Form Recognition Using Artificial Neural Network. 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 420–424, doi: <https://doi.org/10.1109/ICIIS51140.2020.9342638>.

15. Y. Xu, C. Zhang, Q. Zhu and Y. He. A Novel Local Selective Ensemble-based AdaBoost Method for Fault Detection of Industrial Process. 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS), 2020, pp. 1388–1393, doi: <https://doi.org/10.1109/DDCLS49620.2020.9275237>.
16. Crop Yield PredictionDataset. <https://www.kaggle.com/datasets/patelris/crop-yield-predictiondataset>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.