Crop Growth Prediction using Ensemble KNN-LR Model

Attaluri Harshitha[1], Beebi Naseeba[2, *], Narendra Kumar Rao[3], Abbaraju Sai Sathwik[4], Nagendra Panini Challa[5]

[1,2,4,5]School of Computer Science and Engineering (SCOPE), VIT-AP University, India
[3]School of Computing, Mohan Babu University, Tirupathi, India

Abstract

Research in agriculture is expanding. Agriculture in particular relies heavily on earth and environmental factors, such as temperature, humidity, and rainfall, to forecast crops. Crop prediction is a crucial problem in agriculture, and machine learning is an emerging study area in this area. Any grower is curious to know how much of a harvest he can anticipate. In the past, producers had control over the selection of the product to be grown, the monitoring of its development, and the timing of its harvest. Today, however, the agricultural community finds it challenging to carry on because of the sudden shifts in the climate. As a result, machine learning techniques have increasingly replaced traditional prediction methods. These techniques have been employed in this research to determine crop production. It is critical to use effective feature selection techniques to transform the raw data into a dataset that is machine learning compatible in order to guarantee that a particular machine learning (ML) model operates with a high degree of accuracy. The accuracy of the model will increase by reducing redundant data and using only data characteristics that are highly pertinent in determining the model's final output. In order to guarantee that only the most important characteristics are included in the model, it is necessary to use optimal feature selection. Our model will become overly complex if we combine every characteristic from the raw data without first examining their function in the model-building process. Additionally, the time and area complexity of the Machine learning model will grow with the inclusion of new characteristics that have little impact on the model's performance. The findings show that compared to the current classification method, an ensemble technique provides higher prediction accuracy.

*Corresponding author: beebi.naseeba@vitap.ac.in

1. Introduction

Agriculture's complex process of crop forecast has led to the development and testing of numerous models [1]. India is one of the countries that produces the most agricultural goods, but its agriculture output is still very low [2]. So that producers can earn more from the same plot of ground with less labor, productivity needs to be raised. The solution is provided by precision gardening. As the name suggests, precision farming involves adding exact and appropriate amounts of inputs to the crop such as pee, fertilizers, soil, etc. [3,4]. At the right moment to increase output and harvests. Not all precision cultivation techniques produce the greatest outcomes. However, it is crucial that the suggestions made in agriculture are exact and precise because mistakes could result in material loss [5] and financial loss. To develop a reliable and effective algorithm for harvest prediction, numerous studies are being conducted. One such method used in these study projects is assembling. Among the numerous using machine intelligence, this article suggests a system that uses the polling technique to create an effective and exact model, which is currently being used in this area.

Agriculture has long been regarded as India's primary and dominant cultural activity [6]. Due to the fact that

ancient people cultivate their own territory, their requirements have been met [7]. Since the development of new cutting-edge tools and methods in agribusiness is deteriorating, Due to these numerous inventions, people have focused on creating fake and hybrid goods, which can result in a harmful lifestyle [8]. The value of growing goods at the right moment and place is unknown to most people in modern society. These cultivating techniques are also altering the yearly environment at the cost of essential resources like land, water, and oxygen, which leads to food insecurity [9,10]. Each type of learning used in machine learning supervised, uncontrolled, and reinforcement learning has advantages and disadvantages. Supervised training the algorithm takes a set of information used to build a quantitative model that contains both the inputs and the desired outcomes. Unsupervised learning is the process by which a mathematical model is created by an algorithm from a collection of input-only data without identifiers for the intended outputs. With sparse training data and some unlabeled example input, semi-supervised learning algorithms build mathematical models.

It is necessary to use a variety of databases to solve the issue because crop farming relies on both biotic and abiotic variables [11]. The ecosystem's biotic variables are those that result from direct or indirect impacts of living things on other living things (microorganisms, animals, plants, predators, parasites, and pests). Additionally included in this group are human-caused causes (plant protection, fertilization, irrigation, air pollution, soils, and water pollution [12,13] etc.). These variables could result in shape faults, internal defects, and changes in the chemical makeup of the crop yield in addition to other variations in crop output. Both biotic and abiotic factors affect how the ecosystem develops, how plants flourish [14] and how people are affected by it. Three categories of abiotic factors exist: physical, molecular, and other. The recognized physical factors include soil type, topography, rockiness of the soil, climate, and water chemistry, especially salinity [15]. Additionally included are climatic circumstances, ionizing, electromagnetic, ultraviolet, and infrared rays, as well as mechanical movements [17] (vibration, noise). Priority environmental poisons like carbon monoxide, lead, cadmium, nitrogen fertilizers, herbicides, PAHs, nitrogen oxides, fluorine, and its components [18] are just a few examples of the chemical variables. The others are aflatoxins, asbestos, mercury, arsenic, dioxins, and furans [19]. Abiotic elements like water, soil, height, and temperature quality also have an impact on its traits. There are many factors that affect how dirt forms and how valuable it is to cultivation.

In order to increase agricultural production using a range of techniques, this research aims to use machine learning with ensemble and hybrid models for each particular product.

2. Literature Review

We referenced to the study paper as we searched for and read through pertinent articles, papers, and information connected to our project, "Crop Prediction" It is based Using machine learning techniques and data, a systematic evaluation of articles produced classification accuracy for predicting the suitable crop, Accuracy scores of up to 86.0% have been achieved with techniques like Convolutional Neural Network (CNN).

Although these papers impacted on a number of crucial issues, the precision obtained is noticeably lower, and we have made the decision to make use of the helpful information from these papers and make improvements. Last but not least, the study article "Crop Prediction Based on Attributes of the Agricultural Environment Using Different Feature Selection Methods and Classifiers.[1]" nicely matched our goals and provided a road plan for how to approach our project. In order to solve the multiclass classification issue, they tried Nave Bayes, Decision Tree, SVM, KNN, and Random Forest. An intelligent system for crop prediction by soil data, accuracy, precision, and recognition were employed [2]. Ghosh et al. (2023) embarked on a comprehensive study to assess water quality through predictive machine learning. Their research underscored the potential of machine learning models in effectively assessing and classifying water quality. The dataset used for this purpose included parameters like pH, dissolved oxygen, BOD, and TDS. Among the various models they employed, the Random Forest model emerged as the most accurate, achieving a commendable accuracy rate of 78.96%. In contrast, the SVM model lagged behind, registering the lowest accuracy of 68.29% [20].
Alenezi et al. (2021) developed a novel Convolutional Neural Network (CNN) integrated with a block-greedy algorithm to enhance underwater image dehazing. The method addresses color channel attenuation and optimizes local and global pixel values. By employing a unique Markov random field, the approach refines image edges. Performance evaluations, using metrics like UCIQE and UIQM, demonstrated the superiority of this method over existing techniques, resulting in sharper, clearer, and more colorful underwater images [21].

We employed SVM and ensemble model - KNN+Logistic Regression and ANN, and hybrid model that combines CNN-NB, CNN-SVM, CNN-KNN. The Naive Bayes is the method that produces results with the highest precision (99.45%). We obtained accuracy values of 96.67% for ANN, 99% for Decision Tree, 98.36% for KNN, 98% for SVM, 87.81% for CNN, 97.27% for KNN+LR, and 96.59% for the hybrid model (CNN, SVM), 93.64% for hybrid model (CNN, KNN), 92.27% for hybrid model (CNN-NB).

3. Proposed System

3.1. Dataset description

The present dataset was obtained from Kaggle [27] in the format of a Comma Separated Value file, The dataset comprising soil-specific characteristics gathered from Kaggle was used to gather the information for this subject. By improving Indian datasets on topics like soil, temperature, and other topics, the data is generated. The collection has 2200 entries and eight characteristics that describe the factors affecting agricultural output, such as soil pH levels, nitrogen, phosphate, and potassium. The degree of acidity or alkalinity (Ph) is a major factor that influences the amount of soil minerals that are available. The amount of exchangeable aluminum and the activity of soil microbes can both be impacted by PH. The water retention and draining control root penetration. The aforementioned factors are therefore taken into account when picking a harvest for the reasons listed below.

Groundnut, pulse, cotton, vegetables, plantain, rice, sorghum, sugarcane, coriander etc; are among the products taken into account by our algorithm. The training collection displays how many instances of each crop are present. The variables considered were depth, texture, acid, soil color, permeability, discharge, water retention, and erosion. The characteristics of the soil have a significant impact on the crop's capacity to absorb water and nutrients from the earth. Crop growth requires a suitable climate on the earth. The earth acts as a support for the roots. Below Fig 1 illustrates the types of crops and the number of instances for each crop in the dataset.
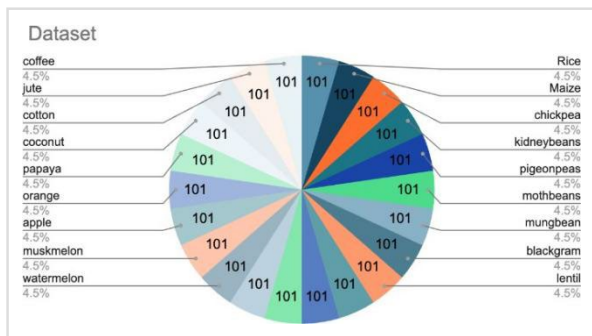

Figure 1. The pie-chart describes about data set.

Additionally, the quantity of precipitation, humidity, and temperature required for specific crops. We use this data to build ANN, SVM, KNN, SGDC, Naive Bayes, Decision tree, ensemble, and mixed versions. The information is divided into 70% for training and 30% for testing and confirmation.

s.no	Title	Models	ACC
1.	Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers [1].	Naïve Bayes Decision Tree SVM KNN Random Forest	70.64% 73.22% 77.50% 83.24% 87.43%
2.	Prediction of crop yield and fertilizer recommendation using machine learning algorithms [2].	Random Forest SVM	86.35% 73.75%

3.2. Methodology

This model is implemented using machine learning techniques which include SVM, SGDC, KNN, Naive Bayes, Decision tree, ANN, Ensemble (KNN & Logistic), and Hybrid model.
The steps below describe how the model works.
Step 1: Import the libraries and load the dataset.
Step 2: Create a dataset with no null values and divide it into the training and assessment sets.
Step 3: Applying the machine learning algorithms.
Step 4: Predict the testing dataset and compute the model's accuracy and confusion matrix.

SVM
A linear model called the support vector machine (SVM) can be applied to classification and regression problems. It works well in both regular and complex situations. The SVM prediction is one of the most effective linear and non-linear binary predictors. SVM works well on small data sets, but because of its higher robustness, performs better on big data sets. Every point in the collection is initially displayed by SVM in an n-dimensional environment and giving a location to each point based on the value of its characteristics in the case of an n-feature dataset.

The SVM method splits data into classes by drawing a line or a hyperplane. The recommendation of the crop performance is classified using the accuracy, confusion matrix, and metric values of precision, recall, f score, and support. In Fig 4 for SVM, the confusion matrix is shown after the validation performance.

SGDC

the confusion matrix is shown after the validation performance. The stochastic gradient descent classifier (SGDC) algorithm is utilized for model optimization. This model can perform well even on large-scale machine learning problems. It is a quantitative optimization iterative method with a maximum number of rounds of 9000. During each iteration, the model trains the numerous variables affecting distinct crops. The method randomly chooses a tiny subset of the training sample and then computes the loss function. The parameters are updated until the smallest loss function is attained. The best fit between the actual crops and the predicted crops is then achieved with the help of these modified parameters. The adjustments made to these factors determine the results and in Fig 2 the confusion matrix is shown after the validation performance.

KNN

A directed machine learning technique called K closest neighbor (KNN) can be used for regression as well as categorization problems. It is non-parametric because it does not make data-driven assertions. KNN can be used to solve categorization and regression issues. In the categorization assignment, the algorithm forecasts the test point's class label based on the class labels of its nearest neighbors. In the regression assignment, the algorithm forecasts the continuous target value of the test point based on the target values of the test point's nearest neighbors. The distance between each data point and the training sample is calculated using the Euclidean distance or Manhattan distance. The model identifies the nearest locations to the data point. The preponderance of the common class among the submissions can be used to decide the crop suggestion. Crops can be recommended using the KNN method based on variables such as weather, humidity, rainfall, and pH values in the earth. A crop recommendation can be calculated using the closest neighbor numbers. In Fig 3 for KNN, the confusion matrix is shown after the validation performance. KNN outperformed the SVM technique in terms of accuracy and speed.

Naïve Bayes

An approach to guided learning used to address classification problems is the Naive Bayes algorithm. The Naive Bayes Classifier is a method for machine learning that produces fast and reliable predictions. It predicts based on an object's likelihood; hence it is a probabilistic classifier. The approach assumes that the features are independent of one another and then calculates the probability of each feature independently for model prediction. It can handle groups of both constant and discontinuous data. It is very expandable and does not contain any unnecessary features. Because the individual premise is consistent with the paradigm, the algorithm exceeds other algorithms. During the training phase, it individually calculates the probability of each crop yield for the given input feature set. This approach categorizes climatic and soil characteristics in order to recommend crops to farmers. In Fig 7 for Naïve Bayes, the confusion matrix is shown after the validation performance. The crop advice would be beneficial in climate change as it reduces the probability of crop failure. As a result, the crop with the greatest likelihood can be recommended.

Decision Tree

A guided machine learning technique that can address both categorization and regression problems is the decision tree classifier. The nodes of the decision tree reflect the characteristics of the crop dataset, and the limbs are the decision criteria that are feature values, and the tree's leaf nodes provide the end class label results. Begin at the base node of the decision tree to forecast the class label. Using the decision criteria and circumstances, this root node feature should now be compared to the other features. In a decision tree, we use an attribute selection measure for selecting the root node and other nodes in the tree. On the values of each attribute, we use entropy, information gain, and the Gini index. These values are placed in the tree, in such a way that better information gain and lower entropy are at the root. After training and testing the model, the crop label prediction is classified according to the identified features. In Fig 8 for Decision Tree, the confusion matrix is shown after the validation performance. The model suggests a crop based on the soil and climate levels determined by the decision tree rules.

ANN

Artificial Neural Networks (ANN) are supervised learning algorithms that are used for categorization and identification. For issue solving, A machine learning programme inspired by the human brain is called an ANN. ANNs are nonlinear statistical models that use intricate input and output relationships to uncover new patterns. The input layer sends the data to the hidden layer, which then sends it to the output layer. This model makes use of two hidden layers. The input level and the concealed level both make use of the ReLU activation function. The activation function's results are sent to the output layer at the concealed layer. The softmax function is suitable because it is a multi-class categorization for the output layer. Backpropagation is used to determine the error based on the output layer outcome. This procedure is done until either the error becomes tiny or the epochs, in this instance 25, are exhausted. In Fig 6 for ANN, the confusion matrix is shown after the validation performance and the Fig 5 depicts the training and testing accuracy. The ANN model provides optimum crop suggestions after it has been taught with crop recommendation influencing characteristics of soil and climate change.

KNN and Logistic Regression Ensemble Model

An ensemble data mining model uses the strength of two or more models to produce predictions and productivity levels that are higher than any single model could produce on its own. One of the most well-known Ensembling methods, Majority Counting, is used in our approach. Any

number of basic learners can be used in the voting method. There should be at least two foundational trainees. The students are selected in such a manner that they complement each other while also being capable of doing so. The greater the rivalry, the better the forecast. However, it is crucial that the students develop their flexibility because when one or a few people make a blunder, the others have a good chance of fixing it. Each pupil fashions a portrait of themselves. The provided training data set is used to train the algorithm. Each model separately predicts the class when new data needs to be classified. Last but not least, the name of the new sample's class is decided upon after polling the majority of students.

The ensemble model, the KNN plus Logistic Regression machine learning model combines the strengths of two distinct models. The KNN and logistic regression are used in the model to increase the model's accuracy or performance. The ensemble model works by training two models on the same dataset, one with a KNN and one with logistic regression. An identical set of input data is used to train each model, but the architectures and weights are different. The results of the two models are then merged to create the final output, which is typically achieved through averaging or weighted averaging.

KNN is a non-parametric model capable of identifying the characteristics most closely connected with successful crop yields. Logistic regression is a parametric model that may predict whether a crop will be successful or not depending on soil and climate factors. It can also identify which characteristics are most significant in determining crop performance. Because KNN and logistic regression are both relatively simple models, the recommendations generated by these algorithms are simple to understand. When compared to a single model, ensemble models that mix KNN and logistic regression can produce more accurate and reliable crop recommendations. By merging the predictions of the two models, the strengths of each model can be utilized and the flaws reduced.

CNN Hybrid Model

Hybrid CNN model with SVM, KNN and Logistic Regression, which will entail picking suitable methods, fine- tuning model settings, and improving model design. To assess the model's success, compare the findings of the hybrid model with those of traditional models such as CNN-SVM. The findings indicate that the suggested hybrid model beats conventional models and has a high level of accuracy in predicting crop yield. Additionally, by analyzing the impact of soil level on the predictions. Fig 11 displays the accuracy analysis comparison for the hybrid model.

4. Results and Discussions

Table 1. Accuracy Comparsions

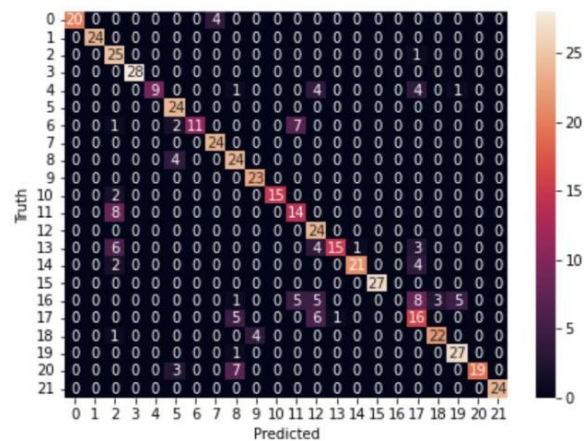| S.NO	Model Name	Accuracy
1	SGDC	79.27%
2	KNN	98.36%
3	SVM	98%
4	ANN	96.67%
5	Naïve Bayes	99.45%
6	Decision Tree	99%
7	Ensemble KNN-LR	97.27%
8	CNN	87.81%
9	CNN-SVM	96.59%
10	CNN-KNN	93.64%
11	CNN-NB	92.27%

Figure 2. A confusion matrix is presented after validation performance for SGDC
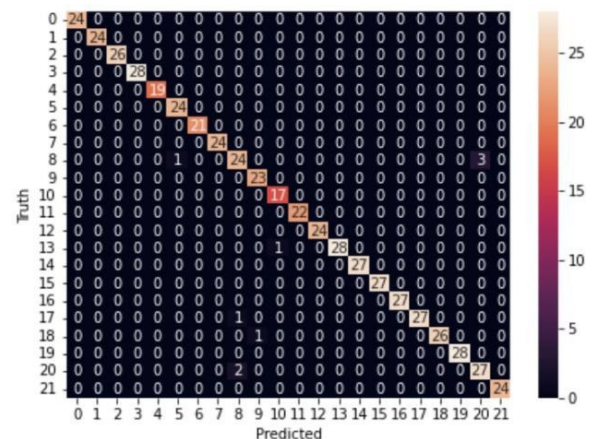

Figure 3. A confusion matrix is presented after validation performance for KNN
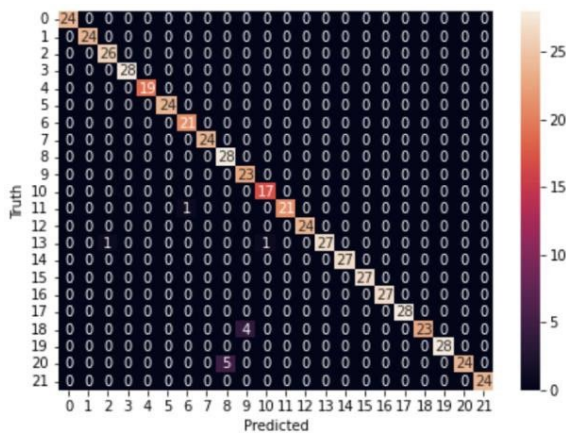
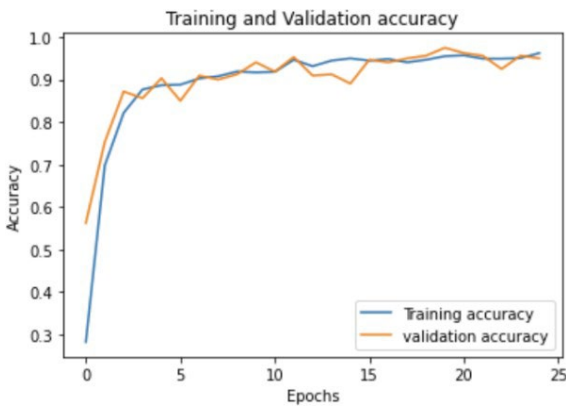**Figure 4.** A confusion matrix is presented after validation performance for SVM

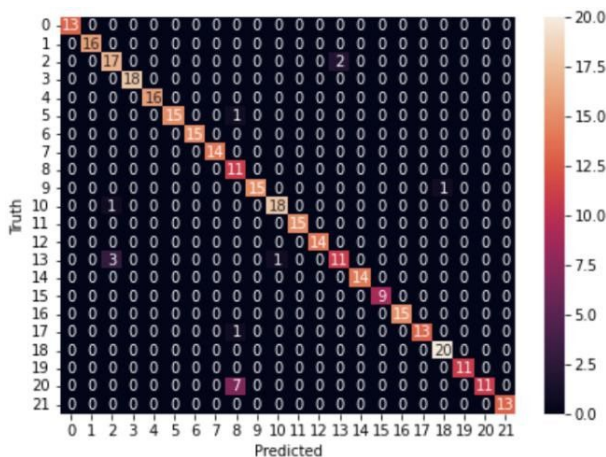
Figure 5. Training vs Validation accuracy graph for ANN

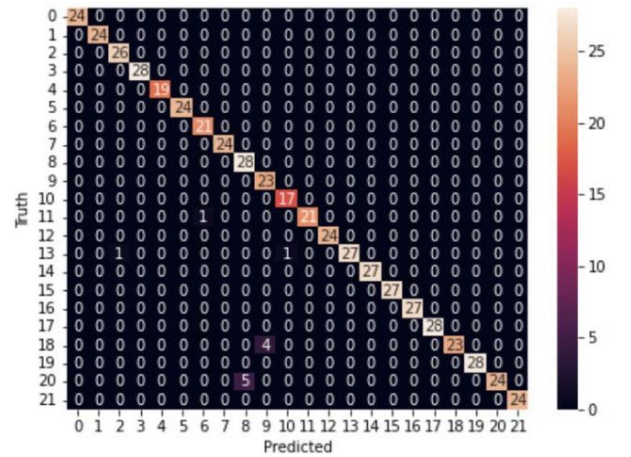**Figure 6.** A confusion matrix is presented after validation performance for ANN

Figure 7. A confusion matrix is presented after validation performance for Naive Bayes
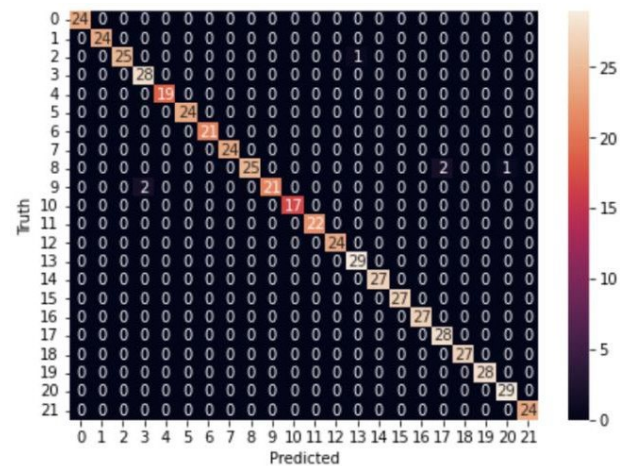

Figure 8. A confusion matrix is presented after validation performance for DECISION TREE
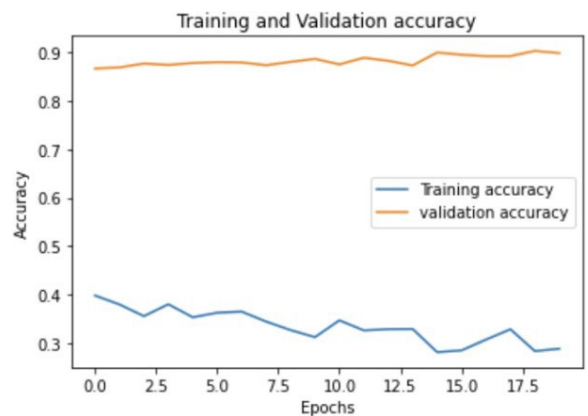

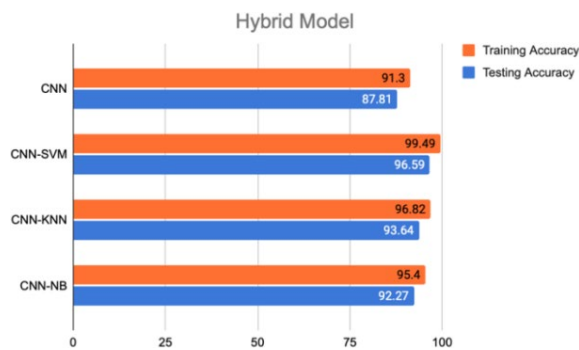**Figure 9.** Training vs Validation accuracy graph for CNN

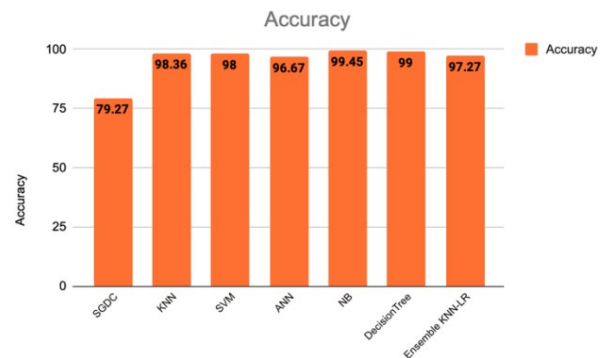**Figure 10.** Comparative accuracy analysis

Figure 11. Comparative accuracy analysis of Hybrid Models

5. Conclusions

The recommendation system is based on the difficulties farmers have in selecting the best crop for production. This model used a variety of machine learning algorithms, producers can select the best crop to cultivate based on the factors influencing farming yield. According to the aforementioned research, hybrid, ensemble, KNN SVM, and ANN are effective for crop forecast with accuracy levels above 96.67% to 99.45%. Naive Bayes and Decision Tree both have high accuracy ratings of 99.45% and 99.01%, respectively. Agriculture might be greatly improved by DL-based crop output prediction, however there are certain issues that need to be resolved. Higher agricultural production is feasible, as shown by the ability to predict crop output dependent on the effective application of algorithms and location.

6. Future Scope

Although crop yield prediction using deep learning (DL) has shown promising results, it is still limited by the need for high-quality data and the difficulty of quantifying some factors, such as weather patterns and soil quality. The accuracy and utility of agricultural production prediction using DL may be improved in a number of future ways. These include integrating additional data sources like weather forecasts and remote sensing data, transferring knowledge from one crop or region to another using transfer learning, optimizing DL models, integrating with precision agriculture methods, and creating user-friendly interfaces. Researchers may concentrate on overcoming these issues as the technology advances and improving the precision and utility of DL-based agricultural production prediction for farmers and other agriculture sector stakeholders.

References

[1] Raja, S.P., Sawicka, B., Stamenkovic, Z. and Mariammal, G., 2022. Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. IEEE Access, 10, pp.23625-23641.

[2] Bondre, Devdatta A., and Santosh Mahagaonkar. "Prediction of crop yield and fertilizer recommendation using machine learning algorithms." International Journal of Engineering Applied Sciences and Technology 4.5 (2019): 371-376.

[3] Rajak, R.K., Pawar, A., Pendke, M., Shinde, P., Rathod, S. and Devare, A., 2017. Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology, 4(12), pp.950-953.

[4] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ''Biotic and abiotic factors influencing on the environment and growth of plants,'' (in Polish), in Proc. Bioróznorodnoss Srodowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne, Puławy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192

[5] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ''Response surface methodology: A retrospective and literature survey,'' J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[6] D. K. Muriithi, ''Application of response surface methodology for optimization of potato tuber yield,'' Amer. J. Theor.Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[7] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ''Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,'' Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183– 188, 2014.

[8] J. R. Oledzki, ''The report on the state of remotesensing in Poland in 2011–2014,'' (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[9] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ''Predicting of blue lupine yields based on the selected climate change scenarios,'' Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.

[10] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ''Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,'' Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[11] N. Chanamarn, K. Tamee, and P. Sittidech, ''Stacking technique for academic achievement prediction,'' in Proc. Int. Workshop Smart Info-Media Syst., 2016, pp. 14–17.

[12] W. Paja, K. Pancerz, and P. Grochowalski, ''Generational feature elimination and some other ranking feature selection methods,'' in Advances in Feature Selection for Data and Pattern Recognition, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97–112.

[13] D. C. Duro, S. E. Franklin, and M. G. Dubé, ''A comparison of pixelbased and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,'' Remote Sens. Environ., vol. 118, pp. 259–272, Mar. 2012.

[14] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, ''Soil classification and suitable crop prediction,'' in Proc. Nat. Conf. Comput. Biol., Commun., Data Anal. 2017, pp. 25–29.

[15] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, ''Deep Gaussian process for crop yield prediction based on remote sensing data,'' in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 4559–4565.

[16] D. A. Reddy, B. Dadore, and A. Watekar, ''Crop recommendation system to maximize crop yield in ramtek region using machine learning,'' Int. J. Sci. Res. Sci. Technol., vol. 6, no. 1, pp. 485–489, Feb. 2019.

[17] N. Rale, R. Solanki, D. Bein, J. Andro-Vasko, and W. Bein, ''Prediction of Crop Cultivation,'' in Proc. 19th Annu. Comput. Commun. Workshop Conf. (CCWC), Las Vegas, NV, USA, 2019, pp. 227–232.

[18] E. Manjula and S. Djodiltachoumy, ''A model for prediction of crop yield,'' Int. J. Comput. Intell. Inform., vol. 6, no. 4, pp. 298–305, 2017.

[19] K. E. Eswari and L. Vinitha, ''Crop yield prediction in Tamil Nadu using Baysian network,'' Int. J. Intell. Adv. Res. Eng. Comput., vol. 6, no. 2, pp. 1571–1576, 2018.

[20] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

[21] Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021, 13, 3470. https://doi.org/10.3390/w13233470