engineering proceedings

MDPI

Proceeding Paper

Optimized Ensemble Learning for Enhanced Crop Recommendation: Leveraging ML for Smarter Agricultural Decision-Making †

Hemalatha Gunasekaran [1,*], Deepa Kanmani Swaminathan [2] and Krishnamoorthi Ramalakshmi[3]

[1] College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, PO. Box: 466, PC: 516, Oman
[2] Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore 641008, India; deepakanmanis@skcet.ac.in
[3] Department of Information Technology, Alliance University, Bengaluru 562106, India;
* Correspondence: hemalatha.ibr@cas.edu.om
† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

Abstract: Agriculture is the backbone of a country and plays a vital role in shaping its economic performance. Factors such as natural disasters, extreme weather changes, pests, and soil quality significantly impact productivity, often leading to economic losses. Accurate predictions in agricultural practices, particularly crop recommendations, can substantially boost productivity and resource management. This research aims to develop a robust crop recommendation system using Ensemble Learning (EL), which integrates multiple machine learning (ML) models for improved performance. The study utilizes two datasets: a real-time dataset available on Kaggle, collected using IoT sensors, and a synthetic dataset generated using CTGAN. These datasets provide crop recommendations for 22 different crops, based on key features like nitrogen, phosphorus, potassium, soil pH, humidity, and rainfall. The performance of various ML models—such as Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Extra Tree Classifier, XGBoost, and Gradient Boost—is compared with that of EL models, including voting, bagging, boosting, and stacking ensemble techniques. The stacking ensemble model achieved the highest accuracy of 99.36% across all ensemble techniques. By further optimizing this model using the Optuna hyper-parameter tuning technique, the accuracy was improved to 99.43%.

Keywords: agriculture; IoT sensors; machine learning models; regression models; ensemble models

Citation: Gunasekaran, H.; Kanmani, D.; Ramalakshmi. Optimized Ensemble Learning for Enhanced Crop Recommendation: Leveraging ML for Smarter Agricultural Decision-Making. *Eng. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 26 November 2024

1. Introduction

Agriculture is a vital asset for any country, contributing to its strength and independence. A nation thrives when it can successfully meet its agricultural needs, becoming self-sufficient and reducing reliance on other countries for daily necessities. The wealth of a country lies in its farming industry and its farmers. However, modern agriculture faces numerous challenges, including global warming, wars, infectious diseases, and pests [1]. To combat these issues AI can be used in crop prediction, weather forecasting, soil health analysis, precision farming, yield prediction, pest control and many more [1,2].

In this research paper, we propose a machine learning (ML) based crop recommendation system. The dataset used in this study is sourced from Kaggle (https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset accessed on 25 August 2024) and was collected using IoT sensors. These sensors measure critical soil parameters including moisture, pH, temperature, humidity, and essential soil

nutrients such as phosphorus (P), nitrogen (N), and potassium (K). Based on these parameters, our system recommends the most suitable crop to achieve improved yield.

2. Literature Review

Dey et al. (2024) [3] divided the dataset into two parts: one containing 11 agricultural plants and the other containing 10 horticultural plants. The authors implemented five distinct machine learning models—SVM, XGBoost, RF, KNN, and DT—on each of the separate datasets rather than on a combined dataset. The XGBoost model achieved an accuracy of 99.09%.

Islam et al. (2023) [4] proposed ML-enabled IOT device to monitor the soil nutrients like N, P, K, pH, temperature, humidity of the soil. The author employed ML based algorithms like catBoost, voting and bagging to predict the recommended crop based on the parameters of the soil. CatBoost obtained the highest accuracy of 97.5%.

Kiruthika et al. (2023) [5] proposed a method based on Improved Distribution-based Chicken Swarm Optimization (IDCSO) with Weight-based Long Short–Term Memory (WLSTM) for crop prediction. The author achieved an accuracy of 95% by employing IDCSO algorithm for feature selection.

Ramzan et al. (2024) [6] implemented ML and EL models on two types of data: real-time data and hybrid data (real-time data and manual data). The author implemented ML algorithms to predict the recommended crop and compared the performance of ML and EL models.

Nikhin et al. (2024) [7] used ML models to predict the crop yield with weather, soil and crop data. The author found Extra Tree Regressor achieved the highest performance among the other ML model followed by Random Forest Regressor and LGBM Regressor.

Elbasi et al. (2023) [8] used fifteen different ML algorithms with a new feature combination scheme. The author achieved 99.59% accuracy using the Bayes Net Algorithm and 99.46% using Naïve Bayes Classifier and Hoeffding Tree algorithm.

S.P. Raja et al. (2022) [9] developed a range of feature selection and classification techniques to predict the yield size of plant cultivations. Their study likely involved identifying key features (such as soil quality, temperature, humidity, and other environmental or agronomic factors) that influence crop yield. By using various machine learning or statistical models, they aimed to classify and predict the expected yield based on these selected features.

Sharma et al. (2024) [10] on crop prediction by demonstrating how different ML models, such as K-Nearest Neighbors (KNN) and deep learning algorithms, can achieve high accuracy in crop selection and disease prediction.

Parween et al. (2021) [11] explored the integration of IoT with ML techniques to create a precise crop prediction system, improving decision-making for farmers through real-time environmental monitoring. The system helped to reduce input costs and boosts productivity by recommending the most appropriate crops based on current soil and weather conditions.

Bakthavatchalam et al. (2022) [12] proposed a machine learning-based system to recommend the best high-yielding crops based on a combination of eight different agricultural attributes. Their aim was to improve precision agriculture using supervised learning algorithms implemented in WEKA. The study evaluated different classification algorithms for crop prediction using multilayer perceptron, rule-based classifier. The performance of the models was evaluated based on accuracy metrics. The results showed that the selected classifiers achieved a high level of prediction accuracy, with a performance rate of 98.2273%.

3. Data Pre-Processing

The dataset used in this study is an IOT sensor dataset available in kaggle. The dataset includes soil nutrients measures like Nitrogen (N), Phosphorous (P) and Potassium (K) and other parameters such as pH of soil, moisture and rainfall with the type of recommended crop as the target variable as shown in Figure 1. The kaggle dataset has 2200 rows and the class distribution of the dataset is shown in Figure 2.

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Figure 1. Sample Dataset.

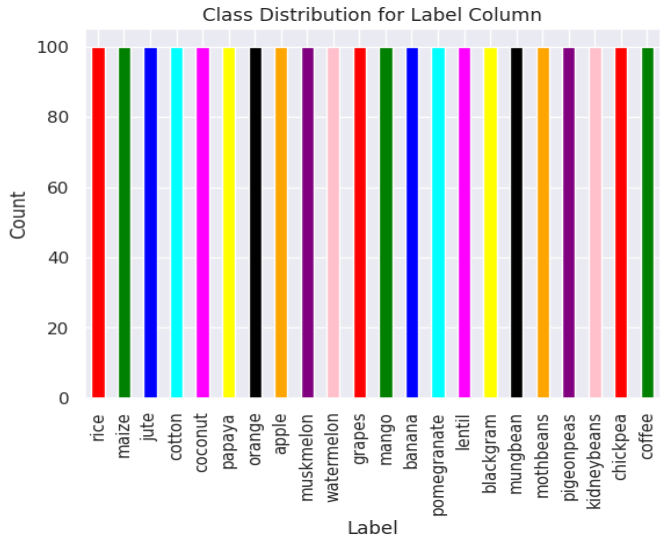

Figure 2. Class Distribution of Original-Dataset.

The kaggle dataset is relatively small, well-balanced, and contains no missing values. However, in real-world scenarios, sensor data can have missing values, noisy data, and errors. To account for this and expand our dataset, we generated approximately 1000 synthetic rows using the CTGAN (Conditional Tabular Generative Adversarial Network) [13] a deep learning model. CTGAN is designed to create synthetic datasets by learning the distribution of the original tabular data, which helps in maintaining the same statistical properties. The CTGAN was trained for 200 epochs with generator learning rate of 0.0002 and discriminator learning rate of 0.0001. The quality of the synthetic data and the kaggle dataset are evaluated, and the graph is given in Figure 3. The synthetic data generated by CTGAN was then concatenated with the original data to create a more robust dataset. This distribution of the combined dataset with 3200 rows is shown in Figure 4. The SMOTE [14] technique is applied to address the issue of imbalanced class data. After implementing SMOTE, each class contains 161 samples each. The details of the dataset are available in Table 1.

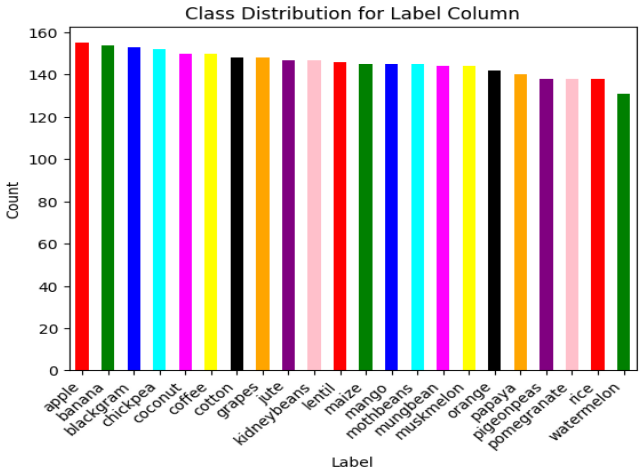

Figure 3. Class Distribution of Original + Synthetic Dataset.

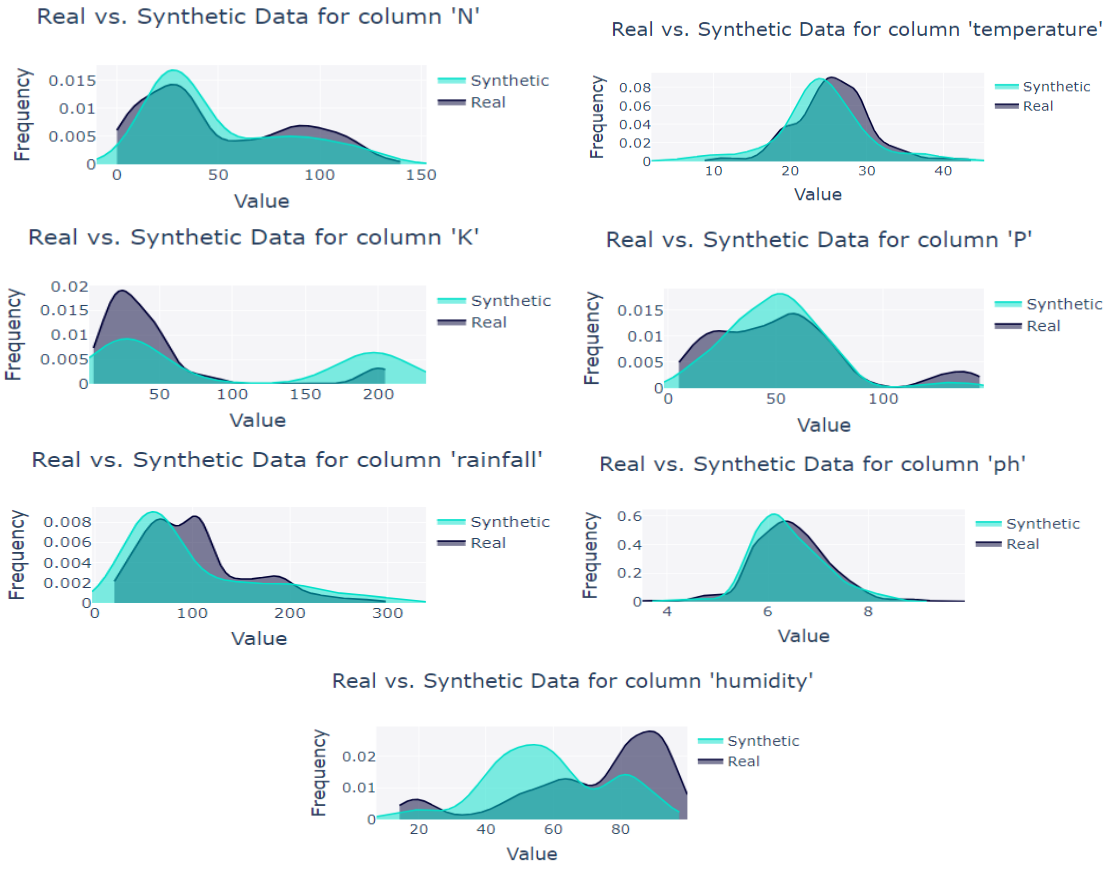

Figure 4. Evaluation of Original and synthetic Dataset using CTGAN.

Table 1. Dataset Details.

	Kaggle Dataset	Kaggle + Synthetic Dataset
No. of rows	2200	3200
No. of Samples per Class	100	161
No. of classes for target variable	22	22

4. Methodology

In this research paper, we have developed a classification model to predict the recommended crop based on seven features: N, P, K, soil pH, temperature, humidity, and rainfall. We have implemented various machine learning (ML) models, including Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), and K-Nearest Neighbour (KNN). Additionally, we have utilized ensemble learning (EL) models such as Random Forest (RF), Extra Trees, XGBoost, Bagging, Gradient Boosting, Voting, and Stacking. Ensemble learning methods provided promising results in many fields [15,16]. The details of the EL model implemented in this research work is given in Table 2.

Table 2. Ensemble Model Parameters.

Model	Base Estimator	No. of Estimator	Meta Classifier
Bagging	Random Forest	100	Nil
Boosting	Gradient	100	Nil
Voting	ExtraTreesClassifier, RandomForestClassifier, XGBClassifier, Decision Tree Classifier	Voting Method: hard	
Stacking	ExtraTreesClassifier, RandomForestClassifier, XGBClassifier, Decision Tree Classifier		LinearRegression

These machine learning (ML) and ensemble learning (EL) models were evaluated on two types of datasets: the original dataset sourced from Kaggle and a concatenated dataset comprising both the kaggle and a synthetic dataset. The synthetic dataset was generated using the CTGAN Generative AI technique. Pre-processing steps were applied to the datasets to normalize the values and convert categorical variables into numerical values, as illustrated in Figure 5. The datasets were then divided into training and testing dataset in the ratio 80% and 20% respectively. The model was created using a training dataset and was tested using testing dataset.

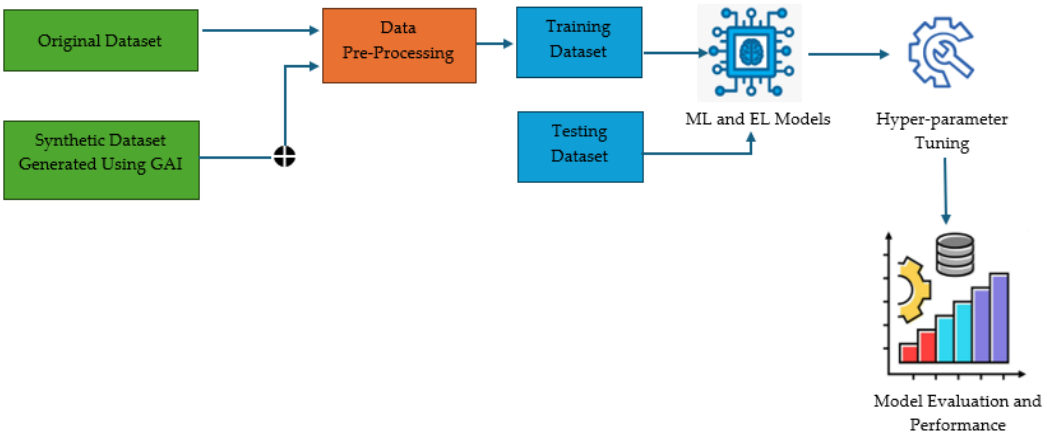

Figure 5. Model of Proposed Approach.

Our findings indicate that the ensemble learning models outperformed the individual machine learning models. However, the synthetic dataset contains more noisy data compared to the kaggle dataset, resulting in lower accuracy. The accuracy, recall and precision of ML and EL models are given in Table 3.

Table 3. Performance Evaluation of ML and EL Models.

Method	Kaggle—Data Set			Synthetic Dataset		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
LR	90.20	87.5	87.621	55.85	55.54	57.42
SVM	89.74	88.18	87.83	46.26	46.10	45.31
DT	97.44	97.27	97.27	69.95	69.75	70.19
NB	98.70	98.64	98.63	57.68	57.49	58.25
K-Neighbor	96.40	95.91	95.90	70.38	70.29	71.55
RF	99.16	99.09	99.09	75.66	74.33	74.50
Extra Tree	98.36	98.18	98.19	75.41	74.33	74.46
XGBoost	98.71	98.64	98.63	73.64	72.64	72.76
Bagging	98.95	98.86	98.86	73.52	72.92	72.95
Gradient Boosting	98.10	97.73	97.74	72.27	70.38	70.60
Voting	98.68	98.64	98.63	73.94	72.78	72.85
Stacking	99.36	99.32	99.32	74.75	73.77	73.87

5. Model Evaluation

The experiment was conducted in Google Colab Pro with Python 3 Google Compute Engine backend with 15 GB GPU RAM and 12 GB of system RAM. The dataset is divided into training and testing in a ratio of 80% and 20% respectively. Seven ML models such as LR, SVM, DT, NB, K-Neighbor, RF and Extra Tree are tested on the kaggle and synthetic dataset. The model performance is improved by implementing EL methods such as bagging, boosting, voting and stacking. The performance of the ML and EL models on the kaggle and synthetic dataset is given in Table 3.

The performance of the models is evaluated based on the metrics such as accuracy, precision and recall as shown below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The performance of stacking ensemble model created with four base learners such as Extra Trees Classifier, Random Forest Classifier, XGB Classifier, Decision Tree Classifier and with one meta learner Logistic Regression obtained an accuracy of 99.36% and it is highest of all the ML and EL models as shown in Figure 6. The proposed stacked ensemble model is compared with the existing models in Table 4. The proposed stacked ensemble model has the highest accuracy of 99.36% when compared to all the other existing models in [3,5,6,9,12]. However, the accuracy of Bayes Net Algorithm discussed in [13] is greater as the author selects only specific features of the dataset for classification. In our proposed method, we utilize the full dataset for prediction. Further, the stacking ensemble model is optimized using optuna an automatic hyper-parameter tuning framework [17, 18]. Optuna finds the optimal combination of hyper-parameters for the ML models. After tuning the stacked ensemble model's accuracy increased to 99.43%. However, the accuracy of the synthetic dataset as given in Table 4 is lower compared to the Kaggle dataset. In this synthetic dataset, the Random Forest (RF) model achieved the highest accuracy of 75.66%, outperforming the other ML and EL models.

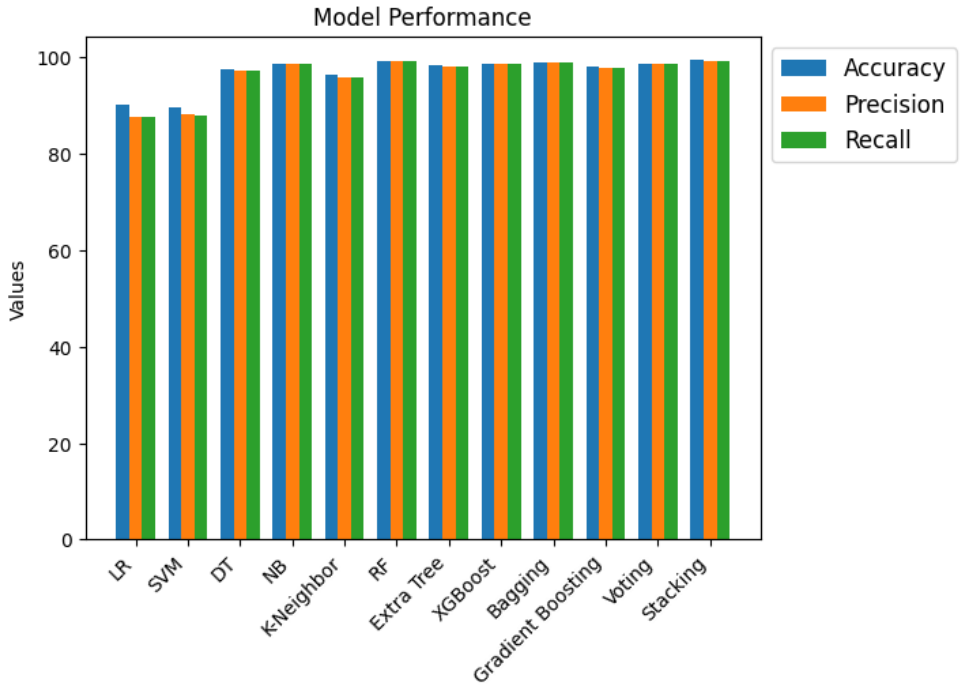

Figure 6. Performance of ML and EL models.

Table 4. Comparison of Proposed and Existing Models.

Reference	Model	Dataset	Accuracy
Elbasi et al. (2023) [13]	Bayes Net Algorithm	Kaggle Dataset with feature selection	99.59%
S.P. Raja et al. (2022) [9]	Bagging	kaggle with MRFE feature selection	97.29%
Biplob et al. (2024) [3]	XGBoost	Kaggle Dataset	99.09%
Ramzan et al. (2023) [6]	KNN	Kaggle Dataset	97.81%
Kiruthika et al. (2023) [5]	IDCSO-WLSTM	Kaggle Dataset	92.68%
Bakthavatchalam et al. (2022) [12]	MLP	Kaggle Dataset	98%
Proposed	Stacking Ensemble with optuna optimization	**Kaggle Dataset**	**99.43%**

6. Conclusions

In this research, the performance of ML models and EL models are compared on both Kaggle and synthetic dataset. We found that EL models perform well on Kaggle Dataset, especially stacked ensemble model created with four base learners such as Extra Trees Classifier, Random Forest Classifier, XGB Classifier, Decision Tree Classifier and with one meta learner Logistic Regression outperformed the other ML and EL models. The performance of the stacked EL model is further improved using optuna optimizer. However, on the synthetic dataset generated using CTGAN, Random Forest achieved the highest accuracy, outperforming both ML and EL models. This study highlights that EL models may not perform well if the dataset contains noisy data, as demonstrated by the lower accuracy on the synthetic dataset.

Author Contributions: Conceptualization, Hemalatha.; methodology, Hemalatha.; software, X.X.; validation, Deepa Kanmani. Ramalakshmi; formal analysis, Deepa Kanmani.; investigation, Deepa Kanmani.; resources, Ramalakshmi.; writing—original draft preparation, Hemalatha.; writing—

review and editing, Deepa Kanmani, Ramalakshmi.; All authors have read and agreed to the published version of the manuscript."

Funding: This research received no external funding

Informed Consent Statement: Not Applicable

Data Availability Statement: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

Conflicts of Interest: The authors declare no conflict of interest

References

1. Thomas van Klompenburg; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709, https://doi.org/10.1016/j.compag.2020.105709.
2. Talaviya, T.; Shah, D.; Patel, N.; Yagnik, H.; Shah, M. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artif. Intell. Agric.* **2020**, *4*, 58–73, ISSN 2589-7217, https://doi.org/10.1016/j.aiia.2020.04.002.
3. Dey, B.; Ferdous, J.; Ahmed, R. Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon* **2024**, *10*, e25112, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e25112.
4. Islam, M.R.; Oliullah, K.; Kabir, M.M.; Alom, M.; Mridha, M.F. Machine learning enabled IoT system for soil nutrients monitoring and crop recommendation. *J. Agric. Food Res.* **2023**, *14*, 100880, ISSN 2666-1543, https://doi.org/10.1016/j.jafr.2023.100880.
5. Kiruthika, S.; Karthika, D. IOT-BASED professional crop recommendation system using a weight-based long-term memory approach. *Meas. Sens.* **2023**, *27*, 100722, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100722.
6. Ramzan, S.; Ghadi, Y.Y.; Aljuaid, H.; Mahmood, A.; Ali, B. An ingenious iot based crop prediction system using ML and EL. *Comput. Mater. Contin.* **2024**, *79*, 183–199. https://doi.org/10.32604/cmc.2024.047603.
7. Nikhil, U.V.; Pandiyan, A.M.; Raja, S.P.; Stamenkovic, Z. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. *Computers* **2024**, *13*, 137. https://doi.org/10.3390/computers13060137.
8. Jhajharia, K.; Mathur, P.; Jain, S.; Nijhawan, S. Crop Yield Prediction using Machine Learning and Deep Learning Techniques. *Procedia Comput. Sci.* **2023**, *218*, 406–417, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.023.
9. Raja, S.P.; Sawicka, B.; Stamenkovic, Z.; Mariammal, G. Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers. *IEEE Access* **2022**, *10*, 23625–23641. https://doi.org/10.1109/ACCESS.2022.3154350.
10. Sharma, K.; Kumar, D. ML- and IoT-Based Crop Prediction System. In *Innovations in Electrical and Electronic Engineering*; Shaw, R.N., Siano, P., Makhilef, S., Ghosh, A., Shimi, S.L., Eds.; ICEEE 2023, Lecture Notes in Electrical Engineering; Springer: Singapore, 2024; Volume 1109. https://doi.org/10.1007/978-981-99-8289-9_44.
11. Parween, S.; Pal, A.; Snigdh, I.; Kumar, V. An IoT and Machine Learning-Based Crop Prediction System for Precision Agriculture. In *Emerging Technologies for Smart Cities*; Bora, P.K., Nandi, S., Laskar, S., Eds.; Lecture Notes in Electrical Engineering; Springer: Singapore; 2021; Volume 765. https://doi.org/10.1007/978-981-16-1550-4_2.
12. Bakthavatchalam, K.; Karthik, B.; Thiruvengadam, V.; Muthal, S.; Jose, D.; Kotecha, K.; Varadarajan, V. IoT Framework for Measurement and Precision Agriculture: Predicting the Crop Using Machine Learning Algorithms. *Technologies* **2022**, *10*, 13. https://doi.org/10.3390/technologies10010013.
13. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. Available online: https://github.com/DAI-Lab/CTGAN (accessed on 26 August 2024).
14. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.
15. Gunasekaran, H.; Gladys, A.; Kanmani, D.; Macedo, R.; Wilfred Blessing, N.R. Brain Stroke Prediction Using Stacked Ensemble Model. *J. Kejuruter.* **2024**, *36*, 1759–1768. https://doi.org/10.17576/jkukm-2024-36(4)-38.
16. Gunasekaran H, Kanmani SD, Ebenezer S, Blessing W, Ramalakshmi K. Detection of Lung and Colon Cancer using Average and Weighted Average Ensemble Models. EAI Endorsed Trans Perv Health Tech [Internet]. 2024 Feb. 5 [cited 2024 Nov. 19];10. Available from: https://publications.eai.eu/index.php/phat/article/view/5017.
17. Elbasi, E.; Zaki, C.; Topcu, A.E.; Abdelbaki, W.; Zreikat, A.I.; Cina, E.; Shdefat, A.; Saker, L. Crop Prediction Model Using Machine Learning Algorithms. *Appl. Sci.* **2023**, *13*, 9288. https://doi.org/10.3390/app13169288.
18. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), Anchorage, AK, USA, 1–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.