Delhi Business Review Vol. 26, No. 2 (July - December 2025)

DELHI BUSINESS REVIEW

An International Journal of SHTR

Journal Homepage: https://www.delhibusinessreview.org/Index.htm
https://www.journalpressindia.com/delhi-business-review

Crop Recommendation System Using LightGBM

Yash Patel[a*], Sunil Kumar[b]

[a] Visiting Faculty, School of IT, AURO University, Surat, India, [b] Associate Professor, School of IT, AURO University, Surat, India.

ARTICLE INFO

**Corresponding Author:*
sunil.kumar@aurouniversity.edu.in

Article history:
Received - 09 June, 2025
Revised - 16 December, 2025
21 December, 2025
23 December, 2025
Accepted - 29 December, 2025

Keywords:
Precision Agriculture,
Crop Cultivation,
Modern Farming Techniques,
Recommendation System,
Ensemble Techniques,
Machine Learning, and
Light Gradient Boosting
Machine (LightGBM).

ABSTRACT

Purpose: The present study is an attempt to study the farmers' challenge's like selecting the best crop for their agricultural site, and measure the impact of precision agricultureas a solution for farmers and most suited to their environment through historical data about soil type and nutrient levels.

Design/Methodology/Approach: To test the research framework and data set, a machine-learning-based ensemble method that recommends crops to grow on an agricultural site using the light gradient boosting machine(LightGBM) algorithm has beenconsidered to achieve higher accuracy and efficiency in recommending a crop at the site.

Findings: Crop recommendationindeed plays a vital role in agriculture for the farmers. The study also found that the LightGBM ensemble machine learning algorithm which produces a series of hypotheses that are then compiled into a final output that maximizes the predictive accuracy of the classification. This study also compared the accuracy, execution time of LightGBM with other algorithms like Adaboost, Gradientboost, Xgboost, Catboost, and found LightGBM far better than the others.

Research Limitations: The study has several limitations. For instance, it engaged the theory of an ensemble machine learning technique to recommend the crop to the farmers. Future, there is a need to expand the techniques byintegrating them with pretrained and LLM models. Furthermore, the sample dataset was selected from an open source of 22 crops with 2200 records.

Managerial Implications: Practically, it brings a focus on the ensemble techniques and theiravailable algorithms with sample dataset implications. The study, thus, showcases the implementation and comparison of an ensemble technique for experimental purposes, knowledge, and competencies to increase with some real-time data in future.

Originality/Value: The study highlighted the importance of ensemble techniques of machine learning supportedby the LightGBM algorithm in the agriculture domain.

DOI: 10.51768/dbr.v26i2.262202504

Introduction

The country's biggest economic asset is agriculture in India. India generates approximately 280 million tons, the second-largest agricultural producer in the world. In 2018, India produced approximately 18% of the total GDP of the world. When compared with conventional farming techniques, traditional farming techniques and procedures have many issues of consistency, cost-effective use, and environmental impact on the resource base. The inability to produce an increased quantity of crops without compromising the quality of the product has been one of the major issues of conventional farming techniques. Agro-based markets around the globe have changed dramatically over the past decade or so. As a result of these changes, many new farming methodologies and innovative methods of crop production have been introduced in India and across many other countries. Precision Agricultural Farming PM is one of these methodologies (Pudumalar et al., 2016). Precision Agriculture (PA) uses computer and IT (technology) tools and techniques, allowing farmers to provide their crops and soils with specific amounts of inputs, ensuring maximum yield and health of the crops and soils. The goal of Precision Agriculture (PA) is to build a decision support system (DSS) with which farmers can effectively manage their entire farm and maximize their input return while preserving the environment. Precision Agriculture can also be referred to as "satellite Agriculture", "Just-in-Time Agriculture", and "site-specific Crop Management (SSCM)". Within the context of Precision Agriculture, crops, fertilizer, and even methods of agricultural production are recommended based on information provided by Precision Agriculture. The area of Crop Recommendation is one of the most important categories within Precision Agriculture. Crop Recommendations can be made based on numerous factors, such as soil conditions, climate, crop type, and market demand, to name just a few. Agricultural research aims to identify specific factors affecting a particular site's agricultural production, rather than relying solely on general recommendations for all regions across the country and world. Precision agriculture offers opportunities to improve crop selection; however, as stated by (Pudumalar et al., 2016), most of the time, Precision Agriculture techniques do not yield accurate and precise results in practice. For those in agriculture, the recommendations provided must be entirely accurate and precise since errors can have serious implications for fiscal and physical production capacity. To this end, many researchers are currently conducting research aimed at developing reliable and efficient models for predicting crop production. A sample methodology employed in this context includes machine learning ensembles, such as those delineated above (Ada Boost, Gradient Boosting, and XGBoost), as well as the LightGBM methodology to produce a more accurate and efficient crop prediction model.

Literature Survey

The report Kulkarni et al. (2018) describes the use of machine-learning ensemble techniques to build a crop forecasting system. The authors have implemented the Random Forest, Naive Bayes, and Linear SVM as individual base learners within the ensemble model. Each classifier independently produces class labels that provide sufficient classification accuracy. Class names assigned by individual base learners are aggregated by a majority voting system. The recommended crops are provided for the Kharif and Rabi seasons. An overall average classification accuracy of 99.91% is achieved through the simultaneous use of independent base learners. The authors of the paper, (Pudumalar et al., 2016), concentrate on using precision farming methods and creating a crop recommendation system to increase crop yields. The authors proposed using an ensemble model that employs the majority voting method combined with Random Trees, CHAID, K-Nearest Neighbor, and Naive Bayes base learners to create a recommendation system for crops suited to the conditions at their location and with maximum accuracy and efficiency.

According to Ujjainia et al. (2021), the incorporation of technology with crop yield forecast techniques has resulted in a significant shift in worldwide output levels. Machine learning has advanced that technology, further improving the situation for farmers and the agriculture industry. To make the agricultural sector competent enough to sustain the expected amount of crop production, the author utilizes the ensemble algorithm for effective prediction.

The requirements and planning necessary for creating a software model to support precision

farming (Babu, 2013) are discussed in detail, beginning with an overview of the fundamentals of precision farming and ending with a software model. The proposed methodology applies Precision Agriculture concepts to small, open farms owned by individual farmers, thus providing some control over the unpredictability of multiple factors affecting the success or failure of a crop. The model's purpose is to provide immediate and real-time support to even the smallest farmer with respect to their smallest plot of crops using the most readily available technologies (e.g., SMS and email). The model has been developed specifically for use in Kerala State, which has the smallest average holding size (as compared to the rest of India), and with very few minor modifications, it could be applied in any region of India. Crop selection and various factors affecting crop selection, including production rate, price in the market, and government policies, are discussed in (Kumar et al., 2015). They propose a Crop Selection Method (CSM) to address the problems of crop selection while improving net yield rates for each crop. This method also provides a suggested sequence of crops to grow during each season by considering parameters such as weather, soil type, crop type, and water density. The projected values of specific parameters will influence the predictive accuracy of CSM.

Methodology

Ensemble Technique and LightGBM Algorithm

Ensemble Learning is a type of Machine Learning in which multiple learners work together to solve Classification Problems. Instead of learning one hypothesis from the data, Ensemble Techniques produce a series of hypotheses that are then compiled into a final output that maximizes the predictive accuracy of the classification. According to Kumar et al. (2015), an Ensemble consists of an arrangement of individual learners called the Base Learners. These learners, referred to as weak learners, are typically created from training data using either Decision Trees, Neural Networks, or other types of ML Algorithms. Figure 1 provides an overview of how Ensemble Techniques Function as described by (Ujjainia et al., 2021).

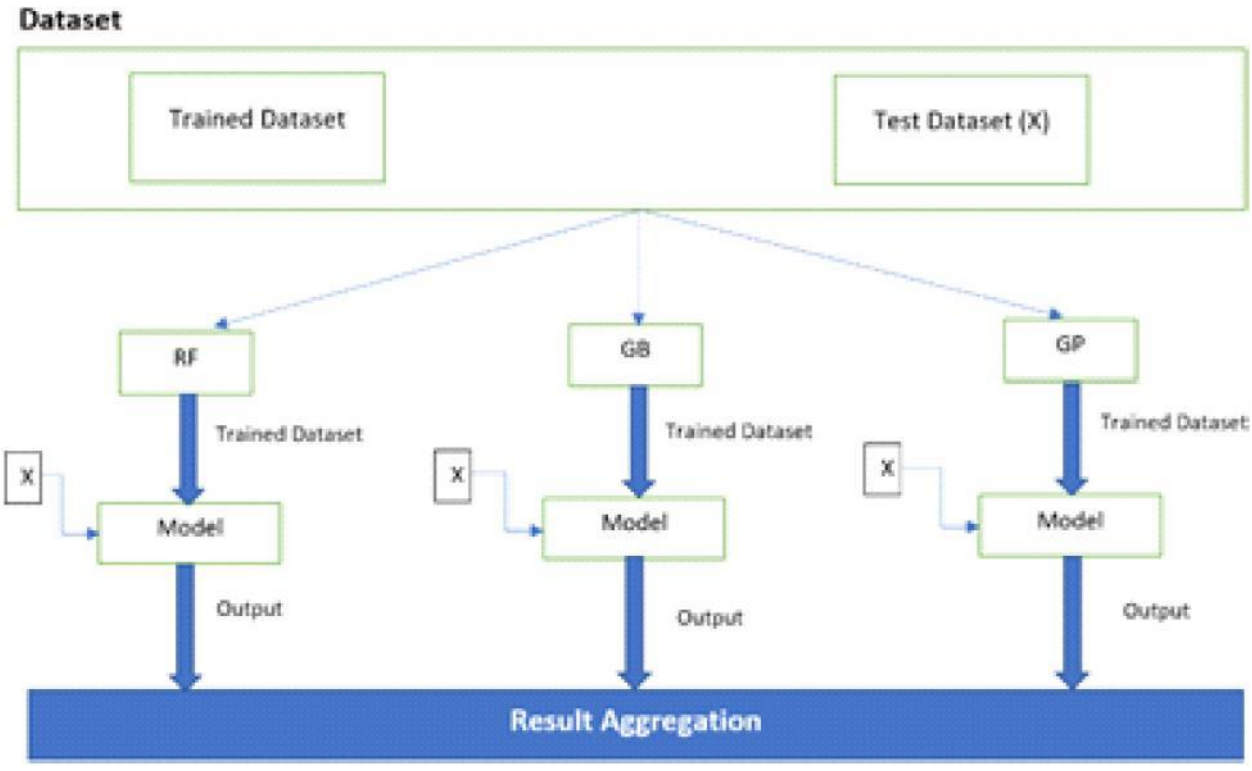

Figure 1: Working Process of the Ensemble Technique

LightGBM is an efficient Gradient Boosting Decision Tree (GBDT) framework that uses Decision Trees to produce fast results with better memory efficiency than other GBDTs. It is considered "Light" because it has been optimized for speed and performance. Light GBM uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to get around the limits of Histogram-based GBDTs. Faliang used GOSS and EFB to develop LightGBM, which grows trees from the leaf nodes rather than level-wise, as do other Boosting Algorithms. An illustration of this process is presented in Figure 2.

Proposed Workflow of the Model

Figure 3 shows the six primary stages of the proposed crop recommendation system: data collection, data preprocessing, data sampling, feature selection, LightGBM classifier, and performance evaluation. The data gathering stage entails acquiring agricultural data and then doing data preprocessing, such as checking for null values and deleting them to enhance crop data quality. Following that, under-sampling is carried out to correct the imbalance of crop samples. In terms of feature selection, it is used to eliminate irrelevant crop feature characteristics in order to improve the efficiency of data training and testing. The best crop is then

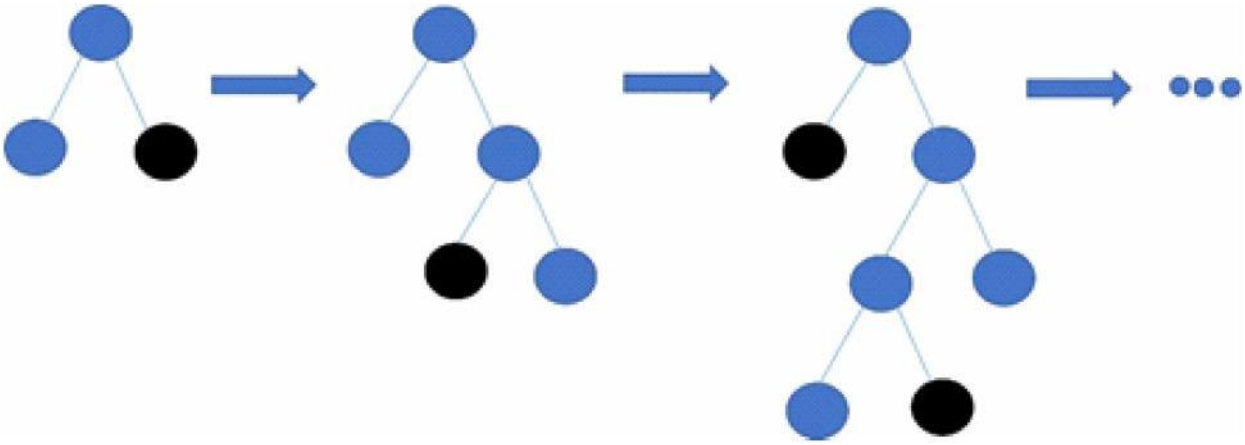

Figure 2: Leaf-wise Tree Growth in LightGBM

Figure 3: Diagram of Proposed Methodology

predicted using a LightGBM-based model based on soil nutrients and climatic parameters. Finally, apply performance evaluation methods such as the confusion matrix, ROC Curve, and Area Under the Curve (AUC), Precision and Recall, and F1-score to analyze the performance of the LightGBM model. A comparison with the XGBoost and Adaboost ensemble algorithms is also presented.

Crop Data Collection

The data set includes soil-specific attributes from sever allocations in India that can be available on the internet. Our model considers 22 crops, including rice, maize, chickpeas, kidney beans, pigeon peas, moth beans, moonbeams, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, musk melon, apple, orange, papaya, coconut, cotton, jute, and coffee, with each crop having 100 records. Nitrogen (N), Phosphorus (P), Potassium (K), Temperature, Humidity, PH, and Rainfall are among the six attributes included in the dataset. Table 1 gives a detailed explanation of these attributes.

Data Preprocessing

There are 2200 records in the collection, each with 8 features. After cleaning the data, the rear eno missing values or irrelevant attributes in the dataset. Then, separate the label column and transform the string labels to integer labels ranging from 0 to 21 using the Label encoding technique, as the dataset comprises 22 different crops.

Data Sampling

In a 70:30 split, divide the complete pre processed data set into two parts: training data and testing data. Ultimately, 1540 records are utilized for training, and the remaining 660 records are used to evaluate model accuracy.

Feature Selection

The random forest (RF) technique is utilized as the base learner for feature selection. The RF is made up of many Decision Trees (DT), each of which is generated using randomly picked samples and features by (Vashishth et al., 2023). When used for feature selection, RF splits all samples into two 'buckets.' The information gain method may be used to compute the significance score of each feature by comparing the classification results on these two buckets after randomly changing the value of a given feature. Entropyis an impurity measure that is used to calculate information gain. Equation 1 is the formula for finding entropy, where S represents sample of given attributes, p+represents the probability of the positive class, and p_represents the probability of the negative class. Then, the formula for finding Information gain is shown in Equation 2, where S(v) represents the samples after the split, and S is the sample before the split

$$E(S) = -(p+) * log2(p+) - (p\_) * log2(p\_) \quad ...(1)$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv| * Entropy(Sv)}{|S|} \quad ...(2)$$

Table 1: Detailed Information About Data Set Attributes

Attributes	Description	Measure Unit
Nitrogen (N)	Ratio of Nitrogen content in soil	Kg/ha
Phosphorus (P)	Ratio of Phosphorus content in soil	Kg/ha
Potassium (K)	Ratio of Potassium content in soil	Kg/ha
Temperature	Temperature	Celsius
Humidity	Relative humidity	Percentage
PH	PH value of the soil	Scale of 0 to 14
Rainfall	Rainfall	Mm

LightGBM Classifier

This multi-classifier model, based on Light GBM, is a recommendation engine that works through multiple iterations of transforming weak leaners into strong leaners. The LightGBM classifier was built using the Gradient-based One-Side Sampling Technique for Crop Recommendations (Hua, 2020). Each data instance has a role to play when computing information gain. For example, data with a greater gradient (untrained data) will provide more information gain than other data. GOSS selects high-quality data with high gradients (for example, individuals who have a higher than defined threshold of gradient and/or those who are among the top percentile of data) while randomly discarding data with lower gradients to preserve the accuracy of the information gain estimation. GOSS-based estimation of information gain will be more accurate than that estimated from uniformly sampled data with the same rate of target samples, particularly due to the large range of the value of information gain. The algorithm behind GOSS is described in Figure 4 (Singh et al., 2019).

Let O be a training data set on a fixed node of a decision tree, and then the variance gain of dividing measure j at a point d for anode is defined in Equation 3 wheren 0 represents $\sum I[xi \in O], n^{i}(d)$ represents $\sum I[xi \in O:xij \leq d]$ and (d)represents $\sum I[xi \in O:xij>d]$.

```
Input:  I: training data, d: iterations
Input:  a: sampling ratio of large gradient data
Input:  b: sampling ratio of small gradient data
Input: loss: loss function, L: weak learner
Models ⟸ {}, fact ⟸ 1 - α / b

topN ⟸ a X len(I), randN ⟸ b X len(I)

for i = 1 to d do
    preds ⟸ models.predict(I)
    g ⟸ loss (I, preds), w ⟸ {1, 1,...}
    sorted ⟸ GetSortedIndices(abs(g))
    topset ⟸ sorted[1: topN]
    randSet ⟸ RandomPick(sorted[topN:len(I)], randN)
    usedSet ⟸ topSet + randSet
    w[randSet] X = fact: Assign weight fact to the small gradient data.
    Newmodel ⟸ L(I[usedSet], - g[usedSet], w[usedSet])
    models.append(newmodel)
```

Figure 4: Gradient-based One-Side Sampling Algorithm

$$V_{j|o}(d) = 1/n_o\left(\frac{\left(\sum_{\{Xi\,\epsilon\,0:Xij\le d\}} gi\right)^2}{n^j_{l|o}(d)} + \frac{\left(\sum_{\{Xi\,\epsilon\,0:Xij>d\}} gi\right)^2}{n^j_{r|o}(d)}\right)$$

...(3)

Gradient One-Sided Sampling, or GOSS, takes advantage of every instance with a larger gradient and performs random sampling on the various instances with small gradients. For each node of the Decision tree, the training data set is represented by the notation O. Equation 4 gives the variance gain of j or the dividing measure at the node's point d. used by (Ke et al., 2017)

Where Ai= {xi € A: xij ≤ d}, Ar= xi € A: xij> d}, Bi= {xi € B: xij ≤ d}, Br=

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i + \frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum_{x_i\in A_r} g_i + \frac{1-a}{b}\sum_{x_i\in B_r} g_i\right)^2}{n^j_r(d)}\right)$$

...(4)

Table 2: Parameters Tuning for LightGBM Classifier

Parameters	Values
boosting_type	Goss (Gradient-based One-Side Sampling)
n_estimators	100
max_depth	1
learning_rate	0.1
subsample_for_bin	200000
subsample	1.0
num_leaves	31

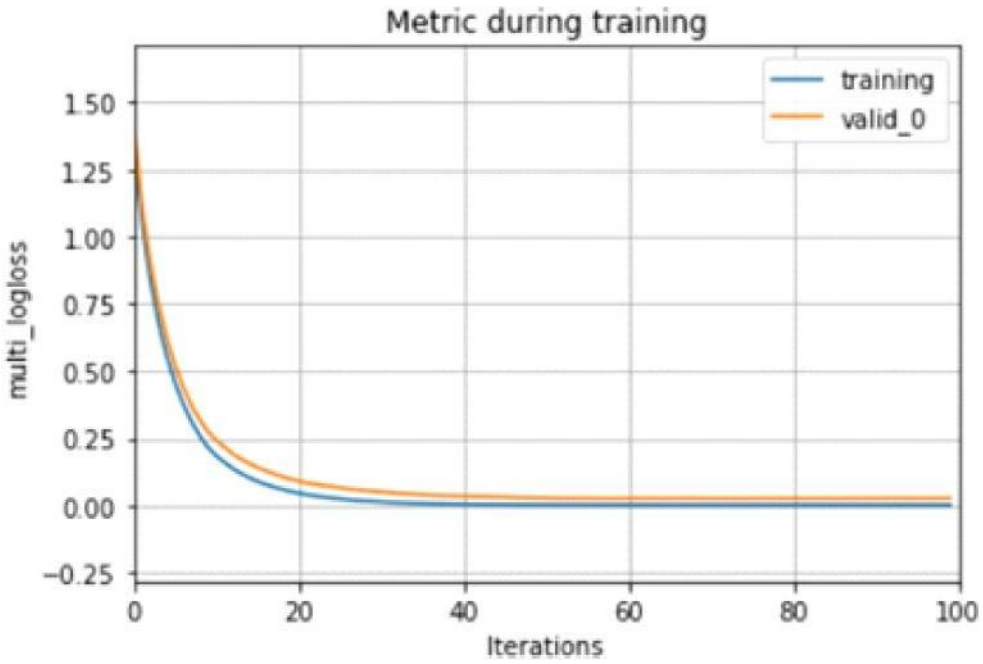

Figure 5: Multi-log loss after 20 Iterations

{xi € B: xij> d} and the coefficient 1-a/ b is used to normalize the sum of gradients over B back to the size of A^c.

Performance Evaluation

The experiments were carried out on a computer equipped with an Intel (R) Core (TM) i5-7200U CPU running at 2.50GHz and 8 GB of memory, where lightgbm and Scikit-learn libraries were used to build the learning model. Table 2 shows the main parameters that are tuned to train the LightGBM classifier. With these parameters, the LightGBM classifier achieved 99 percent accuracy.

The log loss evaluation metric is used to assess the trained model, which is the most important classification metric in terms of probabilities. Figure 5 depicts the training and validating loss after every 20 iterations.

Both training and validating losses decline rapidly within 20 iterations and then remain constant at nearly 0 value. Figure 6 presents an investigation of the effects of crop classification performance on the number of features based on information gain.

Figure 7 shows how well this LightGBM model optimally recommends the best suitable crop based on the given 8 features.

Comparison with Other Gradient Boosting Algorithms

Finally, we examine comparisons of various gradient boosting algorithms on pre-processed data. Figure 8 compares the accuracy of various approaches.

The execution time, accuracy, precision, recall, and F1-score of various ensemble boosting techniques are shown in Tables 3 and 4. It can be seen that the Adaboost classifier takes only 0.89425 seconds to train, but performs poorly in crop classification. XGBoost and Gradient Boost are also accurate and have a high F1-score, but they take too long. The Cat boost algorithm out performs the other approaches with 0.9939 accuracy, but its execution time is slightly longer than that of the LightGBM algorithm. However, when all evaluation parameters, such as accuracy, training time, and other evaluation metrics, are taken into account, LightGBM performs best for the crop recommendation system with 99% accuracy.

Summary

Light gradient boosting machine (LightGBM) ensemble technique using Gradient One-Sided

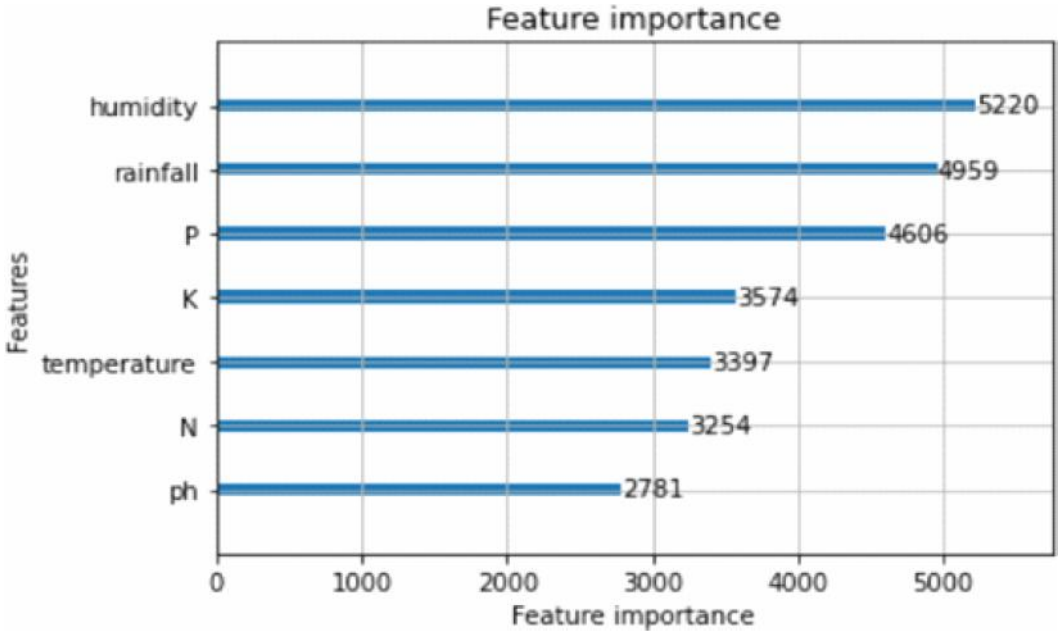

Figure 6: Dominating Features in the LightGBM Classifier

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	28
banana	1.00	1.00	1.00	30
blackgram	1.00	1.00	1.00	31
chickpea	1.00	1.00	1.00	34
coconut	1.00	1.00	1.00	26
coffee	1.00	1.00	1.00	29
cotton	0.97	1.00	0.98	28
grapes	1.00	1.00	1.00	30
jute	0.97	0.94	0.95	31
kidneybeans	1.00	1.00	1.00	26
lentil	1.00	0.95	0.98	22
maize	1.00	0.96	0.98	27
mango	1.00	1.00	1.00	28
mothbeans	0.97	1.00	0.99	36
mungbean	1.00	1.00	1.00	29
muskmelon	1.00	1.00	1.00	30
orange	1.00	1.00	1.00	34
papaya	1.00	1.00	1.00	39
pigeonpeas	1.00	1.00	1.00	28
pomegranate	1.00	1.00	1.00	32
rice	0.95	0.97	0.96	37
watermelon	1.00	1.00	1.00	25
accuracy			0.99	660
macro avg	0.99	0.99	0.99	660
weighted avg	0.99	0.99	0.99	660

Figure 7: Classification Report of LightGBM Classifier

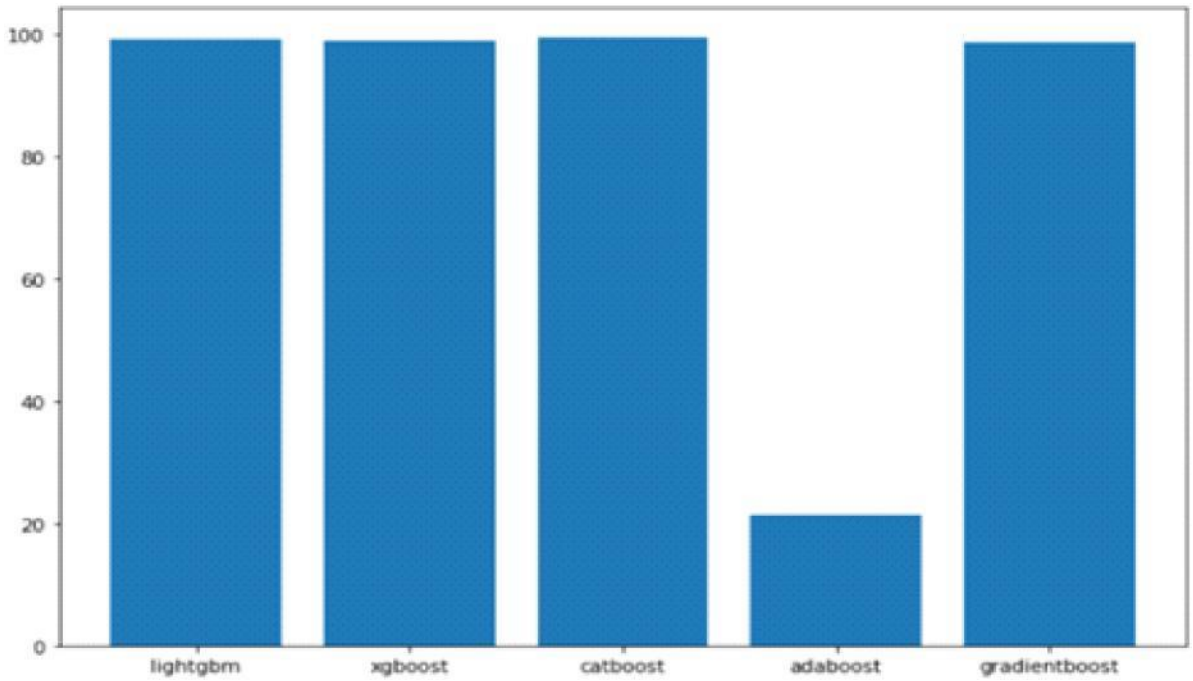

Figure 8: Accuracy of Different Gradient Boosting Approaches

Table 3: Accuracy and Execution Time of Different Gradient Boosting Approaches

Algorithm	Execution time (sec)	Accuracy
Light GBM	1.11918	0.9924
Adaboost	0.89425	0.2121
Gradientboost	9.84128	0.9863
Xgboost	1.17730	0.9909
Catboost	1.93693	0.9939

Table 4: Precision, Recall, and F1-score of Different Gradient Boosting Approaches

Algorithm	Precision	Recall	F1-score
Light GBM	0.9933	0.9920	0.9926
Adaboost	0.160 2	0.2272	0.1697
Gradientboost	0.9876	0.9874	0.9873
Xgboost	0.9919	0.9905	0.9910
Catboost	0.9948	0.9937	0.9942

Sampling (GOSS) works well for the crop recommendation system in this paper. When the proposed LightGBM model is compared to other gradient boosting algorithms such as Catboost, Xgboost, Adaboost, and Gradient boost, the proposed approach significantly out performs in terms of training efficiency and crop classification performance.

References

Babu, S. (2013). Asoftware model for precision agriculture for small and marginal farmers. 2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 352-355. https://doi.org/10.1109/GHTC-SAS.2013.6629944

Hua, Y. (2020). An Efficient Traffic Classification Scheme Using Embedded Feature Selection and Light GBM. 2020 Information Communication Technologies Conference (ICTC), 125-130. https://doi.org/10.1109/ICTC49638.2020.9123302

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS 2017), 1-9.

Kulkarni, N. H., Srinivasan, G. N., Sagar, B. M., & Cauvery, N. K. (2018). Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique. 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 114-119. https://doi.org/10.1109/CSITSS.2018.8768790

Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015). Crop Selection Method to maximize crop yield rate using machine learning technique. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 138-145. https://doi.org/10.1109/ICSTM.2015.7225403

Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T., & Nisha, J. (2016). Crop recommen- dation system for precision agriculture. 2016 Eighth Inter- national Conference on Advanced Computing (ICoAC), 32-36. https://doi.org/10.1109/ICoAC.2017.7951740

Singh, B. P., Kumar, S., & Shekhar, J. (2019). Proceedings of the 1st International Conference on Smart Innovation, Ergonomics and Applied Human Factors (SEAHF) (C. Benavente-Peces, S. Ben Slama, & B. Zafar (Eds.); Vol.

150). Springer International Publishing. https://doi.org/10.1007/978-3-030-22964-1

Ujjainia, S., Gautam, P., & Veenadhari, S. (2021). A Crop Recommendation System to Improve Crop Productivity using Ensemble Technique. International Journal of Innovative Technology and Exploring Engineering, 10(4), 102-105. https://doi.org/10.35940/ijitee.D8507.0210421

Vashishth, T. K., Vikas, Kumar, B., Panwar, R., Kumar, S., & Chaudhary, S. (2023). Exploring the Role of Computer Vision in Human Emotion Recognition: A Systematic Review and Meta-Analysis. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 1071-1077. https://doi.org/10.1109/ICAISS58487.2023.10250614