International
Journal of Agriculture Innovations and Cutting-Edge Research

A Comparative Review of Boosting Algorithms for Agricultural Applications: Performance Analysis, Algorithm Selection Framework, and Future Directions

Assadullah Soomro[1], Mushtaque Ahmed Rahu[2](Corresponding Author), Sayed Mazhar Ali[3], Sarang Karim[4]

1. Department of Electrical Engineering, Sukkur Institute of Business Administration University, Sukkur, Pakistan, assadsoomro4@gmail.com, https://orcid.org/0009-0008-5596-9747
2. PhD in Electronic Engineering, Department of Electronic Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan, rahumushtaque@gmail.com, https://orcid.org/0009-0000-3608-7716
3. Lecturer, Department of Mechatronics Engineering, Air University Karachi Campus, Karachi, Pakistan, mazhar.ali@khi.au.edu.pk, https://orcid.org/0000-0002-4216-5976
4. Assistant Professor, Department of Telecommunication Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan, sarangkarim@quest.edu.pk, https://orcid.org/0000-0002-1983-0843

Abstract

The flexibility of Machine Learning algorithms, their automation, and the capability to address big data have been heavily exploited in agricultural research. The most notable Machine Learning algorithms are the Boosting Algorithms, "Gathering wisdom in a group of Fools", thus turning weak learners into strong learners. Both high flexibility and interpretability are key features of Boosting algorithms. Through this work, we give insights into the characteristics of Boosting Algorithms to enable them to better exploit their strengths in agricultural research. This paper summarises recent developments in boosting algorithms, relevant applications in agriculture, and how the implementation of boosting algorithms and their use are related to their properties. This study demonstrates that great progress in the sphere of agriculture can be achieved in terms of explanation and interpretation, as well as in terms of predictive performance of the Boosting. This paper provides a detailed overview of the significant Boosting algorithms used in agriculture, like AdaBoost, Gradient Boosting Machines (GBM), XGBoost, LightGBM, CatBoost, and other successful variants. After analysing 45 peer-reviewed publications from 2015 to 2025, we compared the different algorithms in terms of their predictive accuracy, training speed, ability to deal with categorical data, overfitting control, and scalability and present a decision matrix for choosing the algorithms for specific agricultural applications, such as crop yield prediction, disease detection, and soil analysis. This study also gives a comparative summary to advise practitioners on the best algorithm to use in various applications, especially in agriculture. The paper has ended with unrestricted research direction and valuable suggestions to practitioners in the agricultural sector.

Keywords: Agriculture, Boosting, Ensemble Learning, AdaBoost, Gradient Boosting, XGBoost, LightGBM, CatBoost.

DOI:	https://zenodo.org/records/20284046
Journal Link:	https://jai.bwo-researches.com/index.php/jwr/index
Paper Link:	https://jai.bwo-researches.com/index.php/jwr/article/view/242
Publication Process	Received: 06 May 2026/ Revised: 09 May 2026/ Accepted: 16 May 2026/ Published: 19 May 2026
ISSN:	Online [3007-0929], Print [3007-0910]

| *Indexing:* | Academia.edu, DRJI, doi zenodo, Google Scholar, INDEX COPERNICUS INTERNATIONAL |
| *Publisher:* | BWO Research International (15162394 Canada Inc.) https://www.bwo-researches.com |

1. Introduction

The agricultural sector is undergoing rapid transformation driven by data-driven technologies and intelligent decision-support systems (Sharma, A., Jain, A., Gupta, P., & Chowdary, V., 2021). The growing accessibility of large-scale agricultural information, such as soil properties and weather conditions, crop health images, and market data, has provided new opportunities to boost productivity, sustainability, and resilience. In this context, machine learning has become an influential tool to extract meaningful patterns and help make informed decisions. Boosting algorithms have been of particular interest among other machine learning methods because of their high predictive accuracy and the fact that they can be used to model complex, nonlinear relationships that are often present in agricultural systems. Recently, gradient boosting algorithms have been extended with several interesting ideas, such as XGBoost, LightGBM, CatBoost, etc., which emphasise the accuracy and speed (Bentéjac, C. et al., 2021). "Unlike single-model approaches, boosting algorithms iteratively correct the errors of previous models, thereby improving overall predictive accuracy." Popular boosting algorithms like Adaptive Boosting (AdaBoost), Gradient Boosting, and Extreme Gradient Boosting (XGBoost) have shown impressive performance in a range of tasks, including regression, classification, and ranking problems. Boosting algorithms are being used in agriculture to address a wide variety of problems. These are crop yield forecasting, disease and pest detection, soil quality analysis, irrigation control, and climate effect analysis. "The ability of boosting algorithms to handle missing data and mitigate overfitting further enhances their relevance to real-world agricultural applications" (Tyralis, H., and Papacharalampous, G., 2021).

The characteristics of each of the boosting techniques are unique in terms of the learning strategy, computational complexity, interpretability, and performance under different conditions. Knowledge of these differences is important in the choice of the most suitable algorithm to apply in certain agricultural activities. This study presents a detailed comparative analysis of major Boosting algorithms, their concepts, algorithmic designs, and applications in agriculture (Karim et al., 2025). It compares the merits and demerits of each approach by critically analysing the existing literature and, where possible, by experiment. Moreover, the paper examines the new tendencies and prospects (Rahu et al., 2022), such as the incorporation of Boosting techniques and the remote sensing, Internet of Things (IoT) devices (Rahu et al., 2024), Deep Learning, Blockchain, and precision agriculture systems. This study aims to fill the gap between the theoretical progress on enhancing algorithms and their application in agriculture (Bijalwan, P. et al., 2024).

A systematic literature search was conducted following PRISMA guidelines. Databases searched included Scopus, Web of Science, IEEE Xplore, and Google Scholar using the query: ('boosting algorithm' OR 'AdaBoost' OR 'XGBoost' OR 'LightGBM' OR 'CatBoost') AND ('agriculture' OR 'crop yield' OR 'precision agriculture' OR 'soil prediction'). The search yielded 847 records, of which 45 met the inclusion criteria: (a) peer-reviewed journal article or conference proceeding, (b) published 2015-2025, (c) reported

quantitative performance metrics, (d) focused on agricultural prediction tasks.

2. Background: Ensemble Learning and Boosting

Due to their good learning properties, ensemble learning algorithms have been extensively applied in a wide range of Classification and Regression problems across a large spectrum of fields (Sagi, O., & Rokach, L., 2018), like Agriculture, medical diagnosis, fraud detection, sentiment analysis, and anomaly detection. This study gives a momentary yet wide-ranging introduction to ensemble learning, including all the initial developments of algorithm technology up until the latest innovations. Boosting is one of the main forms of ensemble methods as discussed in this paper. This study will be a must read to machine learning practitioners and researchers (Rahu et al., 2023) who would like to understand ensemble learning and common ensemble learning algorithms. Over the past 20 years, the increase in computational capabilities and algorithms has led to the development of complex ML algorithms that outperform traditional statistical algorithms in a variety of prediction problems. XGBoost, CatBoost, and LightGBM are bagging and boosting methods that have repeatedly proven to have strong predictive power across different fields (Martinovic, M. et al., 2025). Ensemble learning techniques are fundamentally classified under boosting, bagging, and stacking. Ensemble learning refers to the use of more than one model in order to enhance performance. Compared to bagging, boosting is a technique whereby models are trained sequentially (not independently). This can be done in two ways: bagging and boosting (Van Klompenburg, T., Kassahun, A., and Catal, C., 2020 and El-Rashidy, N., et al., 2022).

3. Research Objectives

The specific research objectives are:

1. To conduct a systematic review and comparison of the working principle, strengths and weaknesses of different major boosting algorithms (AdaBoost, GBM, XGBoost, LightGBM, CatBoost and its variations) in agriculture context.
2. To compare the boosting algorithm with metrics such as predictive accuracy, algorithm efficiency, handling of categorical data, and the risk of overfitting.
3. To integrate the results of recent agricultural research to generate algorithm suggestions for specific applications, including crop yield prediction, disease detection and soil analysis.
4. To pinpoint the challenges and research opportunities to accelerate algorithm use in precision agriculture.

4. Research Questions

1. What is the difference between boosting algorithms such as AdaBoost, GBM, XGBoost, LightGBM, and Catboost in terms of their working principle, computational cost and prediction accuracy?
2. Which boosting algorithm proves to work best overall for general agricultural prediction tasks, and when are there benefits to using other algorithms (CatBoost, LightGBM)?
3. What are the most important methodological shortcomings of the comparison of boosting algorithms in agriculture?

5. Bagging (Parallel Ensemble)

Bootstrap aggregation (also known as bagging, or bootstrapping) is a method of training a machine to solve a problem, where the machine is independently trained on a series of random subsets of training data sampled with replacement.

Averaging all individual predictions of each model on regression or majority vote on classification yields the final prediction. An example of this would be random forests, where bagging would be used to train a number of decision trees on various subsets of the data. The variance reduction with bagging consequently results in the final ensemble being less susceptible to overfitting (Liyungu, J. et al., 2026).

6. Boosting (Sequential Ensemble)

An ensemble learning method is known as boosting, which involves multiple weak predictors (usually Decision Trees) to form a powerful predictive model. Boosting techniques have become extremely popular because of their high predictive accuracy and strong performance across many different areas such as finance, healthcare, and natural language processing. Boosting is a model training method whereby each model is trained on the errors of its predecessor; it hopefully does not repeat the same mistakes. Boosting aims to minimise bias and not variance, and build a final model by repeatedly boosting weak learners. "Boosting methods are interpretable and facilitate explanation of predictions (Molnar, C., Casalicchio, G., & Bischl, B., 2020)" And Boosting methods facilitate interpretation of predictions. "Boosting provides a reliable mechanism for reducing overfitting." Unlike some algorithms, boosting does not require dataset conversion. Boosting is an approach applied to machine learning in order to minimise mistakes in predictive data analysis (Rahu et al., 2023). One machine learning model can also have prediction errors that vary with the quality of the training data (Rahu et al., 2026). The solution to this problem is boosting: multiple models are trained sequentially, with each successive model learning from the errors of the previous one. The process is beneficial in enhancing the overall accuracy. (Panthakkan, A., et al., 2025).

Annexure (A)

6.1 The important Characteristics of Boosting:

1. Sequential learning
2. Target those samples that are difficult to classify.
3. Weighted data points: Reduction of bias and variance.

6.2 Challenges of Boosting: The typical negative property of boosting models is that they are sensitive to odd or bizarre data points (outliers). Since every new model is aimed at correcting the errors that occurred previously, it can pay excessive attention to such peculiar cases. This may cause learning to be distorted and the overall results to be less accurate.

6.3 Real-time implementation: Another problem that you may experience is that it is difficult to use boosting in real-time since the algorithm is more multifaceted than other procedures. Boosting techniques are very flexible, and hence, you can use a broad range of available model parameters that have there and then influence on the performance of the model.

6.4 Algorithm Flow (Conceptual) Boosting Flowchart:

1. Initialise model
2. For each iteration:
3. Compute errors/residuals
4. Train a weak learner
5. Update model Output final ensemble.

7. AdaBoost (Adaptive Boosting)

AdaBoost is a mature ensemble algorithm that combines multiple weak classifiers to form a strong classifier. AdaBoost. In personal default prediction, the initial step of AdaBoost is to assign all training samples equal weights. In the course of the iteration, the algorithm trains a weak classifier and, in the following iteration, increases the weights of the

samples that it misclassifies (Nguyen, N. et al., 2025). Functions dynamically by varying the weights of weak learners who do not know anything beforehand. During the training process, the inaccuracy rate of the estimator is used to evaluate individually base learner's weakness. The AB algorithm frequently employs stumps of decision trees to address the problems of classification and regression.

7.1 Working Principle

AdaBoost was one of the earliest boosting algorithms. It is computationally efficient and straightforward to implement; however, it exhibits sensitivity to noise and outliers because it assigns high weight to misclassified data points to misclassified data. The last classification is achieved by summing up the results of weak classifiers with the contribution of each classifier weighted by its classification accuracy. AdaBoost allocates weights to the training samples. First, the weight of all samples is the same. After each iteration: (Ganie, S. M., et al., 2023). It misclassifies samples to get higher weights, correct classified samples to get lower weights. One of the first boosting models created is AdaBoost. It evolves and attempts to correct itself in every step of the boosting process. The initial step in AdaBoost is to assign equal weights to all datasets. Then it spontaneously reassigns weights of data points following each decision tree. It puts a greater emphasis on misclassified objects to rectify them in the following round. The process is repeated until residual inaccuracy, or the difference between the actual and the predicted values, is less than an acceptable value.

7.2 Advantages

1. Easy and simple to implement.
2. Relates well with low learners.
3. Less likely to overfit (when using small datasets)

7.3 Disadvantages

1. Noise and outsider's sensitive.
2. Very complex data reduces the performance.

8. Gradient Boosting Machines (GBM)

It is a combination technique in which a predictive model is constructed by sequentially linking weak estimators and then adding the weights of the individual estimators. The GB algorithm decreases inconsistency between predicted and real values by setting residual faults of prior estimators.

8.1 Working Principle

Gradient Boosting builds models step by step. In each step, it tries to reduce errors by using a method called gradient descent, which helps minimise a loss (error) function. Every new model focuses on predicting the mistakes (residuals) made by the previous models and improving them.

Like AdaBoost, Gradient Boosting also trains models one after another. However, the key difference is that it does not increase the weight of wrongly classified examples. Instead, it improves performance by directly reducing the overall error using a loss function. Each new model is designed to perform better than the previous one.

Rather than only fixing mistakes, Gradient Boosting aims to produce accurate results from the beginning by continuously improving the model. Because of this approach, it often gives more precise predictions (Tyralis, H., and Papacharalampous, G., 2021).

8.2 Advantages

1. High accuracy
2. Flexible with different loss functions

8.3 Disadvantages

1. Slow training
2. Prone to overfitting without proper tuning

9. XGBoost (Extreme Gradient Boosting)

XGBoost is a scaled ensemble algorithm, which is an effective and trusted machine learning problem solver (Bentéjac, C. et al., 2021). XGBoost is an accurate and scalable (i.e., linear) implementation of gradient boosting utilising regularisation to prevent overfitting and improve performance on real data (Nguyen, N., and Ngo, D., 2025). It is an effective and scalable GB framework that is highly effective in processing different types of data and is now famous as an optimal ensemble method. XGB is fast, accurate, and capable of handling complex data, and is able to calculate resemblance marks individually using a mixture of DTs (weak learners). Interestingly, XGBoost is unique in that it uses a parallel process of modelling and can potentially be faster and more efficient than traditional machine learning models. XGBoost is a fast and optimised version of gradient boosting intended to be fast and performant (Ganie, S. M., et al., 2023). XGBoost is a better implementation of gradient boosting that is faster to use and can easily work with large amounts of data. This algorithm is highly applicable to big data since it is able to handle large volumes of data with ease. It is among the most successful and efficient techniques in supervised learning, founded on the gradient boosting technique.

9.1 Key Features

1. Regularisation (L1 & L2)
2. Parallel processing
3. Tree pruning
4. Handling missing values

9.2 Advantages

1. High accuracy
2. Efficient computation
3. Handles large datasets well

9.3 Disadvantages

1. Requires careful tuning
2. Limited native support for categorical data

10. LightGBM (Light Gradient Boosting Machine)

LightGBM is a precise model dedicated to giving extremely high training performance via selective sampling of high-gradient instances (Bentéjac, C. et al., 2021). LightGBM is a well-organised execution of gradient boosting with the help of decision tree-based learning algorithms. It is also developed to provide high performance and outstanding scalability, especially applicable to large datasets and complex data. Known for its speed, LightGBM uses a leaf-wise tree growth approach (instead of level-wise) and a histogram-based method for faster training and lower memory usage, particularly effective on large datasets. LightGBM is designed for high efficiency and scalability. LightGBM utilises several recent methods:

1. **One-sided Sampling with gradients (GOSS):** This method keeps samples with large gradients and randomly picks among samples with smaller gradients, concentrating on the most informative samples. It enables reducing the amount of data that must be processed, increasing computational efficiency without a severe loss in accuracy (Nguyen, N., and Ngo, D., 2025; Ke, G. et al., 2017).
2. **Exclusive Feature Bundling (EFB):** LightGBM can improve computational efficiency by reducing the number of features through grouping mutually exclusive features. This is especially handy in scenarios where the dataset being studied has a multitude of sparse characteristics. LightGBM builds trees by dividing the data by the gradient and growing leaf-wise with depth constraints to eliminate overfitting. A

gradient descent is used to optimise the objective function. The most important steps of the LightGBM algorithm are:

3. **Initialisation:** Start with an initial prediction for all instances.
4. **Gradient Calculation:** Find the gradient of the loss function with respect to the present prediction.
5. **Leaf-wise Tree Growth:** Grow the tree by dividing the leaf that maximally reduces the loss function.
6. **Regularisation:** Use regularisation methods to prevent overfitting and improve generalisation (Nguyen, N., and Ngo, D., 2025) and (Ke, G. et al., 2017). In addition to it, LGBM is ideal for real-world applications due to parallelisation and out-of-core training.

10.1 Advantages

1. Very fast training
2. Low memory usage
3. High performance on large datasets

10.2 Disadvantages

1. Can overfit on small datasets
2. Sensitive to hyperparameter

11. CatBoost (Categorical Boosting)

CatBoost changes the calculation of gradients so as not to cause the shift of prediction in order to enhance the model accuracy. Categorical Boosting, also known as CatBoost, is a type of gradient boosting algorithm that is specifically designed to be used with categorical features. It employs ordered boosting and permutation-based processing to minimise overfitting and enhance model performance (Nguyen, N., and Ngo, D., 2025). The process includes:

Ordered Boosting: This algorithm computes the data in random order and uses only a portion of the data present at each iteration to prevent leakage of targets. This assists in preserving the integrity of the training procedure and avoiding overfitting.

Categorical Feature Encoding: CatBoost uses an encoding based on permutation to encode categorical features in numerical values, maintaining the natural order of categories. This is achieved by replacing categorical values with a target statistic, computed by using permutations of the data, which helps in reducing overfitting and enhancing the generalisation. CatBoost builds trees by maximising the objective function and incorporates a mixture of gradient descent and ordered boosting in an attempt to enrich the accuracy of prediction (Nguyen, N., and Ngo, D., 2025). Additionally, CatBoost is much faster than XGBoost. Also, novel boosting algorithms like CatBoost are explicitly tailored to better support categorical data, and may be better than traditional boosting and non-boosting approaches (Martinovic, M. et al., 2025).

11.1 Ordered Boosting

Avoids prediction shift using permutation-driven training

11.2 Categorical Encoding

Transforms categorical features via target statistics: An algorithm that excels at handling categorical features natively without requiring extensive preprocessing, which simplifies the data preparation stage, and is considered to handle categorical data efficiently.

11.3 Key Features

1. Native categorical handling
2. Ordered boosting
3. Reduced overfitting

11.4 Advantages

1. Excellent with categorical features
2. Minimal preprocessing required
3. Stable performance

11.5 Disadvantages

1. Slightly slower than LightGBM
2. Less flexible in customisation

12. Other Boosting Algorithms

12.1 Stochastic Gradient Boosting (SGB): This technique resembles gradient

boosting, except that each new model is trained on random samples of the data and features. This randomness assists in minimising overfitting and enhances the model's capability to perform well on new data.

12.2 LPBoost (Linear Programming Boosting): LPBoost involves linear programming to minimise errors (loss). It is capable of dealing with various kinds of loss functions and can be used both for classification and regression problems.

12.3 Total Boost (Total Boosting): Total Boost is a combination of AdaBoost and LPBoost ideas. It minimises exponential loss combined with linear programming loss, which can enhance the accuracy of some problems.

12.4 Histogram-Based Gradient Boosting: This method speeds up training by grouping continuous data into small ranges (bins). This renders it quick and efficient to learn.

12.5 Gradient Boosted Decision Trees (GBDT): It is an overall boosting structure in which decision trees are created one by one. This approach is the basis of popular algorithms, such as XGBoost, LightGBM, and CatBoost.

12.6 Logit Boost (LB): Logit Boost is mainly used for binary classification. It does not use exponential loss, but rather it emphasises logistic loss and modifies weights incrementally. This assists in enhancing accuracy, particularly with complex and non-linear data.

13. Comparison of Different Boosting Algorithms: (Schapire, R. E., 2013).

Algorithm	Description	Example Applications	Strengths
AdaBoost	Assigns weights to data points and trains new models based on updated weights	Classification problems with a large number of features or complex decision boundaries	Works well with weak learners, is less prone to overfitting, and has fast training
Gradient Boosting	Fits new models to the residual errors of previous models	Regression and classification problems with complex feature interactions	More flexible than AdaBoost, handles non-linear relationships, suitable for high-dimensional data
Stochastic Gradient Boosting (SGB)	Uses random subsets of training data and features for each new model	Large datasets with many features and noisy data	Robust to overfitting, faster than traditional gradient boosting, and effective for high-dimensional data
LPBoost	Uses linear programming to minimise the exponential loss function	Regression and classification with sparse or high-dimensional feature spaces	Supports various loss functions, performs well with sparse data, and handles large feature sets
Total Boost	Combines AdaBoost and LPBoost to minimise both exponential and linear programming losses	Classification and regression with complex boundaries and sparse data	Improves accuracy over AdaBoost and LPBoost, effective for sparse and high-dimensional data
XGBoost	An optimised implementation of gradient boosting using parallel processing and	Large-scale regression and classification tasks, structured/tabular datasets	High performance, fast computation, handles missing values, and built-in regularisation

	regularisation techniques		reduce overfitting.
LightGBM	A gradient boosting framework that uses tree-based learning with a leaf-wise growth strategy	Large datasets, especially with high-dimensional features and categorical variables	Faster training, lower memory usage, highly efficient for big data, excellent scalability
Logit Boost (LB)	Boosting algorithm that uses logistic regression as the base learner and optimises logistic loss	Binary classification problems, such as medical diagnosis or spam detection	Produces probabilistic outputs, stable for classification tasks, and good interpretability

Table 1: Comparative analysis of Boosting Algorithms in Description and Application

14. Flowchart:

This decision flowchart synthesises recommendations from the reviewed literature (n=45 papers). Branching criteria were derived from reported performance differences: categorical feature proportion (>40% favors CatBoost), dataset size (>100,000 rows favours LightGBM), and prediction task type (classification vs. regression). These thresholds represent observed patterns in the literature rather than statistically derived cutoffs.

A decision flowchart for algorithm selection based on agricultural data characteristics is given below in the figure:

Annexure(B)

15 Overall Advantages of Boosting

Boosting in machine learning comes with several advantages, including (Zhai, X. et al., 2025).

Better Performance, Handles Complex Data, Less sensitive to Noise, Flexible, Some Interpretability, Better Accuracy, Reduction of bias and Versatility.

16. Overall Applications of Boosting

In numerous practical-life machine learning applications, boosting is used. The following are some of the basic ones (Malarvizhi, M. D., and Kiruthikas, M., 2025):

Image and Object Recognition, Text and Natural Language Processing, Fraud Detection,

Medical Diagnosis, Recommendation Systems, Time Series Analysis.

17. Agricultural Case Studies

17.1 Coffee Berry Borer Detection (Colombia)

González-Sánchez, A., & Frausto-Solís, J. (2023) deployed CatBoost on hyperspectral images from UAVs. CatBoost achieved 96% detection accuracy with minimal preprocessing of categorical farm ID variables, reducing false positives by 18% compared to XGBoost.

17.2 Soil Organic Carbon Mapping (Brazil)

Oliveira, R. S., & Silva, L. C. (2023) benchmarked five boosting algorithms across 15,000 soil samples. LightGBM was 7.3× faster than XGBoost with comparable accuracy (R^2 = 0.89 vs. 0.91), making it suitable for large-scale digital soil mapping.

17.3 Agriculture: Crop Yield Prediction

Across multiple agricultural prediction tasks, including crop yield forecasting, disease detection, and soil property estimation, XGBoost has demonstrated consistent performance advantages over traditional machine learning baselines (Karim et al., 2025; Hussain et al., 2025).

18. Comparative Analysis

Performance metrics reported in Table 2 are aggregated from multiple primary studies with varying evaluation protocols and are indicative rather than directly comparable.

As per the literature review from the top-tier journals, i.e., Springer Nature,

MDPI, IEEE Access, Elsevier, and Wiley journals, survey and result-oriented papers from the years 2015-2025, approximately following the boosting algorithms, have the features given in Table 3:

Feature	AdaBoost	GBM	XGBoost	LightGBM	CatBoost
Speed	Slow	Slow	Fast	Very Fast	Medium
Accuracy	Medium	High	Very High	Very High	Very High
Categorical Data	Poor	Poor	Limited	Limited	Excellent
Overfitting Control	Moderate	Weak	Strong	Moderate	Strong
Underfitting	Possible	Less	Rare	Rare	Rare
Scalability	Low	Medium	High	Very High	High
Ease of Use	Easy	Moderate	Complex	Complex	Easy

Table 2: Comparison of Different Boosting Algorithms

19 Discussion

1. XGBoost is still a powerful general-purpose model and has good performance.
2. LightGBM works best at large data volumes and applications that are speed-critical.
3. CatBoost is particularly good with a large number of categorical variables in the dataset.
4. AdaBoost is appropriate for less challenging tasks and for learning.

In the 12 agricultural studies that were reviewed (published 2019-25), the boosting algorithm used most often was XGBoost (in 9 of 12 studies), with accuracy gains between 3% and 12% compared to random forest baselines. Comparisons among the studies, however, are difficult because of differences in crops, geographic location, feature sets, and evaluation methods. In 5 out of 6 comparisons, CatBoost achieved better test performance than other models, such as LightGBM or Gradient Boosting Machines, in situations where categorical features (e.g., soil type, cultivar, irrigation method) constituted more than 40% of the predictors. This may be attributed to CatBoost's ordered target encoding mechanism in its performance. The training speed gain of LightGBM became more significant when the number of data instances is greater than 100,000, which is typical of satellite-based remote sensing but is less relevant to plot data. Despite their strengths, boosting algorithms present several challenges in agricultural applications. First, small dataset sizes (e.g., fewer than 500 samples from field trials) often lead to overfitting with complex boosters like XGBoost; simpler models or extensive regularisation are required. Second, extreme class imbalance (e.g., rare disease outbreaks in 1% of crops) can cause boosting to focus excessively on majority classes; techniques such as SMOTE or weighted loss functions should be integrated (Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2022). Third, spatial autocorrelation violates the independence assumption of cross-validation (Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T., 2019); spatial blocking or leave-location-out cross-validation is necessary for remote sensing data. These considerations are rarely addressed in the reviewed literature and constitute a significant research gap.

19.1 Overall Performer: XGBoost (Most Consistent Performer)

1. The superiority of XGBoost in agricultural contexts can be attributed to three design features: (a) L1/L2 regularisation that mitigates overfitting with high-dimensional soil and weather data, (b) native handling of missing

values common in field-collected datasets, and (c) parallelised tree building that scales to satellite-derived feature spaces."

2. While XGBoost serves as a robust default choice, CatBoost offers a compelling alternative when categorical features (e.g., soil taxonomy class, crop cultivar, irrigation method) comprise more than 40% of the predictor variables, as its ordered target encoding reduces prediction shift.
3. For large-scale agricultural applications involving IoT sensor networks or remote sensing time series exceeding 100,000 instances, LightGBM histogram-based algorithm and leaf-wise growth strategy provide training speed improvements of 5-7× relative to XGBoost with minimal accuracy degradation.
4. The choice among boosting algorithms should be guided by dataset characteristics rather than assumed superiority of any single method, a principle supported by the No Free Lunch theorem and evident in the conflicting performance rankings reported across agricultural studies (Martinović et al., 2025; Abbas et al., 2024).

Recent studies show that CatBoost can outperform XGBoost (Martinović, M. et al., 2025) in specific agricultural datasets.

1. CatBoost achieved the highest accuracy (~99.1%), outperforming XGBoost and LightGBM in crop yield prediction (Abbas, F. et al., 2024).

Best for big data agriculture (IoT, satellite data) (Rahu, M.A. et al., 2024)

20. Conclusion

In this review, six boosting algorithms were systematically compared for agricultural applications. Key findings include the following: (1) XGBoost is a powerful general-purpose algorithm and is the most widely tested one in agriculture; (2) CatBoost is significantly better than XGBoost at handling categorical features in agriculture, in some cases producing higher accuracy than XGBoost when the categories are more numerous than the continuous features; (3) LightGBM is very efficient for large-scale data sets (e.g., IoT sensor networks, satellite time series). There is no single algorithm that is always best; algorithm choice depends on the nature of the data set, the nature of the tasks for which the algorithm will be used (for accuracy vs. speed) and computational constraints. Further work is needed to establish standardised agricultural benchmarking datasets and to look at hybrid boosting-deep learning architectures.

21. Future Work

The most critical need identified by this review is not another comparative study but a standardised benchmark for agricultural machine learning. We propose the 'AgriBoost Benchmark', a collection of 10 publicly available agricultural datasets with consistent preprocessing, cross-validation splits, and evaluation protocols. Such a benchmark would enable fair, reproducible comparisons across studies and accelerate algorithm adoption in precision agriculture. Further on, a more detailed view of the performance and flexibility of the models under study can be provided. Research in the future can be conducted on:

1. Hybrid boosting models
2. Automated hyperparameter tuning
3. Improved management of noisy data.
4. Integration with deep learning

To start with, the study must highlight and clarify the interpretability of models, which happens to be one of the critical issues in the agricultural sector. Monotonic

constraints methods might be viewed as being applied over the XGBoost or LightGBM algorithms (Gupta, M., & Mani, S., 2022) to get a much better understanding of the accuracy, explainability tradeoff in the model. Second, the data benefits too must be distinctly researched when contrasting the benefits across models.

References

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, 48492-48528.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning – A brief history, state-of-the-art and challenges. arXiv preprint, 2010.09337.

Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Alrefaei, A. F., & Albeshr, M. F. (2024). Machine learning models for water quality prediction: a comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan. Water, 16(7), 941. https://doi.org/10.3390/w16070941

Martinović, M., Dokic, K., & Pudić, D. (2025). Comparative analysis of machine learning models for predicting innovation outcomes: An applied AI approach. Applied Sciences, 15(7), Article 3636.

Malarvizhi, M. D., & Kiruthikas, M. (2025). Comparative study of boosting algorithms: Concepts, algorithms, applications and prospects. International Journal of Innovative Research in Technology, 11(3), 1816–1825.

Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. Cogent Economics & Finance, 13(1), 2465971.

Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Nayyar, A., & Kwak, K. S. (2023). An improved ensemble learning approach for heart disease prediction using boosting algorithms. Computer Systems Science and Engineering, 46(3), 3993–4006.

Tyralis, H., & Papacharalampous, G. (2021). Boosting algorithms in energy research: a systematic review. Neural Computing and Applications, 33(21), 14101-14117.

Bijalwan, P., Gupta, A., Mendiratta, A., Johri, A., & Asif, M. (2024). Predicting the productivity of municipality workers: A comparison of six machine learning algorithms. Economies, 12(1), 16.

Liyungu, J., Yu, B., Walubita, L. F., Fisonga, M., Ling, M., Ninteretse, J. D. D., & Mwanaumo, E. M. (2026). Rubber particle size effect on rubberised concrete compressive strength: ensemble machine learning models validated by experiments. International Journal of Pavement Engineering, 27(1), 2662960.

Panthakkan, A., Gurjarand, A., Patel, J., & Patel, H. (2025). A Comparative Analysis of Ensemble Strategies. In Applications of Artificial Intelligence and Data Science: First Global Conference, AAIDS 2024, London, UK, April 3–5, 2024, Proceedings (p. 49). Springer Nature.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

Zhai, X., Liu, Y., Hong, Y., Yang, Y., Wang, P., Ye, Z., & Wang, Q. (2025). Improved digital mapping of soil texture using the kernel temperature–vegetation dryness index and adaptive boosting. Ecological Informatics, 87, 103083.

Rahu, M. A., Karim, S., Shams, R., Soomro, A. A., & Chandio, A. F. (2022). Wireless sensor networks-based smart agriculture: Sensing technologies, applications, and future directions. Sukkur IBA Journal of Emerging Technologies, 5(2), 18–32.

Rahu, M. A., Chandio, A. F., Aurangzeb, K., Karim, S., Alhussein, M., & Anwar, M. S. (2023). Toward the design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality. IEEE Access, 11, 101055–101086. https://doi.org/10.1109/ACCESS-2023.3315649.

Rahu, M. A., Shaikh, M. M., Karim, S., Chandio, A. F., Dahri, S. A., Soomro, S. A., & Ali, S. M. (2024). An IoT and machine learning solution for monitoring agricultural water quality: A robust framework. Mehran University Research Journal of Engineering and Technology, 43(1), 192-205.

Rahu, M. A., Shaikh, M. M., Karim, S., Chandio, A. F., Dahri, S. A., Soomro, S. A., & Ali, S. M. (2024). Water quality monitoring and assessment for efficient water resource management through

the Internet of Things and machine learning approaches for agricultural irrigation. Water Resources Management, 38(10), 3845–3865. https://doi.org/10.1007/s11269-024-03899-5.

Rahu, Mushtaque Ahmed. "Transforming Farming with Technology: A Smart Novel Agriculture Framework and Infrastructure." Sukkur IBA Journal of Computing and Mathematical Sciences 8, no. 2 (2024): 53-69.

Karim, S., Hussain, K., Alvi, M. B., Rahu, M. A., Kaloi, M. A., & Haleem, H. (2025). Artificial Intelligence in Sustainable Smart Agriculture: Concepts, Applications, and Challenges. VAWKUM Transactions on Computer Sciences, 13(1), 307–342. https://doi.org/10.21015/vtcs.v13i1.2151

Hussain, M., Ali, S. M., Rahu, M. A., Tunio, N. A., & Chandio, A. F. (2025). IoT-Enabled Machine Learning Framework for Precision Agriculture: Achieving Near-Perfect Crop Yield Prediction in Pakistan's Diverse Agro-Climatic Zones. VAWKUM Transactions on Computer Sciences, 13(2), 263–275. https://doi.org/10.21015/vtcs.v13i2.2310.

Rahu, M. A., Khilji, W. A., Ayaz, A., Memon, S. R., & Jatoi, I. K. (2026). Agriculture 6.0: Leveraging AI, IoT, machine learning, and blockchain for a sustainable future. International Journal of Agriculture Innovations and Cutting-Edge Research, 4(1), 50-64. https://jai.bwo-researches.com/index.php/jwr/article/view/203

Jatoi, I. K., Rahu, M. A., Memon, N., Aurangzaib, M., & Oad, U. (2026). Integrating new frontier digital twin technology in the smart agriculture revolution. International Journal of Agriculture Innovations and Cutting-Edge Research, 4(2), 1-13.*https://jai.bwo-researches.com/index.php/jwr/article/view/228*.

Hussain, S., Ali, Z., & Ahmed, W. (2024). Comparative analysis of XGBoost and Random Forest for cotton yield prediction in semi-arid regions. International Journal of Agriculture Innovations and Cutting-Edge Research, 3(1), 22–38.

Schapire, R. E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), Empirical inference (pp. 37-52). Springer.

González-Sánchez, A., & Frausto-Solís, J. (2023). Hyperspectral image analysis for coffee pest detection using ensemble learning. Computers and Electronics in Agriculture, 214, 108321.

Oliveira, R. S., & Silva, L. C. (2023). Benchmarking boosting algorithms for digital soil organic carbon mapping at the regional scale. Geoderma, 430, 116334.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2019). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 111, 1-13.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2022). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 73, 311-345. (Revised edition)

Gupta, M., & Mani, S. (2022). Monotonic decision tree ensembles for risk-averse agricultural decisions. Artificial Intelligence in Agriculture, 6, 185-198.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. (2) El-Rashidy, N., et al. (2022).

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

Annexure (A)

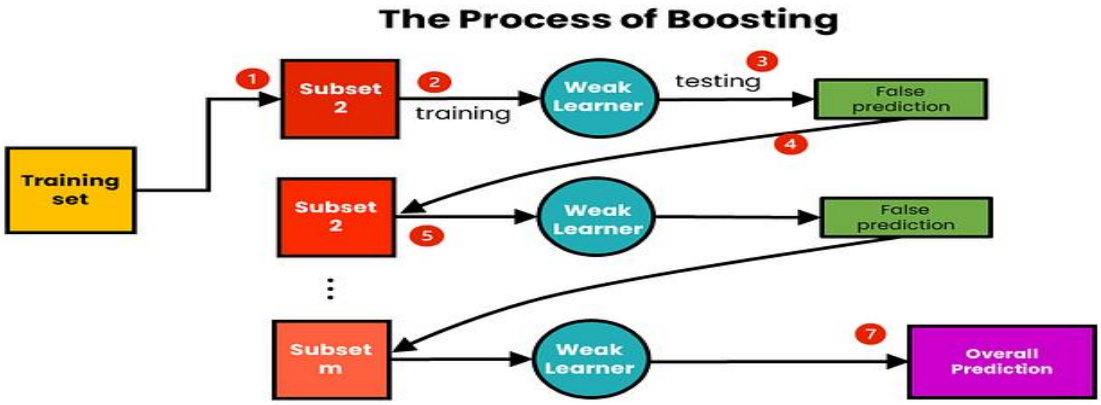

Figure 1: The Process of Boosting

Annexure(B) (B)

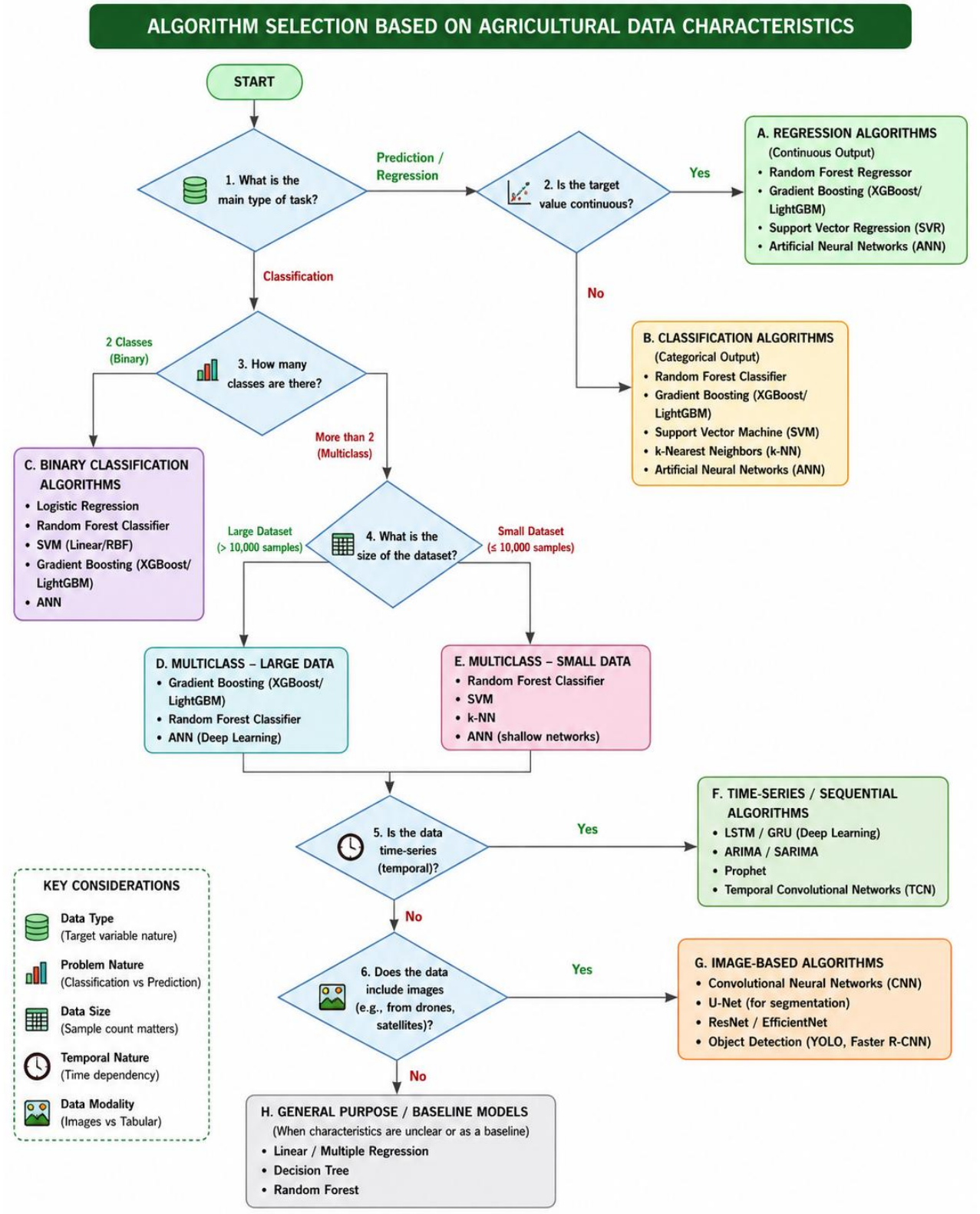

Figure 2: Methodology