Proceedings of 8th International Conference on Mechanical, Industrial and Energy Engineering 2024
02-04 January, 2025, Khulna, BANGLADESH
SciEn Conference Series: Engineering Vol. 3, 2025, pp 85-91 https://doi.org/10.38032/scse.2025.3.19

Multivariate Jute Forecasting Using Ensemble Learning for Supply Chain Optimization

Mohammad Morshed[*], *Faiyaz Bin Mahmud, Md Jawad Bin Rouf, Mohammad Mynul Islam Mahin*

Department of Mechanical and Production Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh

ABSTRACT

Reliable forecasts of crop production and yield are critical to improving food security and optimizing agricultural supply chains. Through utilizing machine learning (ML) techniques, this study seeks to optimize Bangladesh's jute supply chain through multivariate forecasting. It focuses on evaluating six ensemble learning models in order to forecast jute yield and production. Algorithms including Random Forest (RF), Gradient Boosting (GB), Category Boosting (CatBoost), Adaptive Boosting (AdaBoost), Light Gradient Boosting (LightGBM), and Extreme Gradient Boosting (XGBoost) were employed and their performance was thoroughly assessed to forecast yield rate and production on the basis of historical data. Performance was evaluated utilizing mean squared error (MSE) and R^2 scores to determine the best model. With the lowest MSE of 0.0820 and the highest R^2 of 0.9199, LightGBM was found to be the best-performing model, showcasing its superior accuracy in identifying intricate patterns and complex dependencies in the dataset. The findings demonstrate how LightGBM may lessen inefficiencies and lessen supply chain instability in the jute industry, which has historically been plagued by erratic weather patterns and shifting consumer demand. This study applies ensemble learning techniques specifically to jute forecasting and is one of the first to combine yield rate and production in a single multivariate framework. It is anticipated that the improved predictive capability of the model will help stakeholders such as farmers, traders, and policymakers to enhance decision-making, reduce waste, and promote better resource allocation across the supply chain. Moreover, the study's established ensemble learning framework's scalable nature provides a means of extending its use to other agricultural sectors in Bangladesh. This study also advances data-driven approaches to crop management and supply chain optimization, which boost sustainability and profitability in the jute sector and beyond.

Keywords: Agriculture, Machine Learning, Supply Chain Management

1. Introduction

The jute industry in Bangladesh, often referred to as the "Golden Fiber," holds immense historical and economic significance due to its pivotal role in the nation's economy and its strong presence in global markets. Jute is one of Bangladesh's primary agricultural exports, accounting for about 3.5% of the country's agricultural GDP [1], and employing nearly 25 million people across its supply chain from cultivation and processing to export [2]. As the world's second-largest producer of jute, Bangladesh contributes approximately 42% of global jute production [2]. The country annually exports over $1 billion worth of jute and jute-related products, including raw jute, jute yarn, hessian cloth, and sacks [3]. These exports not only provide vital foreign exchange but also support the livelihoods of millions of farmers and workers. However, despite its critical role, the jute supply chain faces challenges, such as unpredictable demand, inefficiencies in production planning, and climate variability. These issues often result in supply bottlenecks, price volatility, and financial losses for farmers and exporters [4]. One of the central issues facing the jute industry in Bangladesh is the lack of accurate forecasting models that can handle the complexity of agricultural supply chains. Traditional forecasting methods, often based on historical data or simple regression techniques, struggle to account for the dynamic and multivariate nature of the factors influencing jute production, such as fluctuating weather patterns, soil quality, and market demand [2]. For instance, in 2023, adverse weather conditions, particularly irregular monsoon rains, contributed to a 10% decline in jute production, leading to shortages in supply and an increase in prices [5]. In such scenarios, the absence of a reliable forecasting model not only causes economic distress for farmers and producers, but also destabilizes the entire supply chain, resulting in losses for exporters and market inefficiencies. Moreover, the growing demand for sustainable, eco-friendly products worldwide adds pressure to the jute industry, highlighting the need for an optimized and resilient supply chain.

A systematic literature review was conducted through the utilization of protocols depicted in **Table 1**.

[*] Corresponding Author Email Address: m.morshed.mpe@aust.edu

Table 1 Research protocol utilized for systematic literature review.

Research protocol	Brief description
Databases	Google Scholar
Language	English
Timeline	2014 to 2024
Search Keywords	"forecasting" AND "jute" AND "supply chain" OR "staple foods" AND "Bangladesh" OR "machine learning" OR "optimization"
Inclusion criteria	Reports or articles highlighting the agricultural forecasting and supply chain optimization
Exclusion criteria	Reports or articles (i) that are not indexed in Google Scholar; (ii) published in languages other than English; (iii) that lack relevance to the specific RQs.

The use of machine learning (ML) techniques has gained substantial traction in agriculture, especially for forecasting crop yields and optimizing supply chains. In this context, ML algorithms can assess complex and large datasets to identify patterns and relationships that traditional statistical models may overlook. For example, Dahikar & Rode (2014) explored the use of Artificial Neural Networks (ANN) for predicting crop yields based on parameters such as soil pH, nitrogen levels, temperature, and rainfall. Their findings showed that ANN could offer more accurate predictions than conventional methods, especially in handling non-linear and multivariate data. The method's ability to learn from complex patterns in large datasets without requiring explicit programming makes it particularly effective in agricultural predictions. However, the challenge with such models is the need for well-structured and clean data, which is often available in technologically advanced regions, but less so in developing countries like Bangladesh. In Bangladesh, research into the ML applications in agriculture is still emerging [6].

Shakoor et al. (2017) applied supervised learning methods to improve agricultural production forecasts for major crops, including jute. Their study utilized algorithms such as Decision Tree Learning (ID3) and K-Nearest Neighbors Regression (KNN) to predict crop output based on historical data from the Yearbook of Agricultural Statistics and the Bangladesh Agricultural Research Council Information System. The ID3 algorithm was used for classification by constructing decision trees based on entropy and information gain, while KNN regression employed a distance-based approach to predict crop yields based on similar historical data points. Their findings demonstrated how ML can improve decision-making, moving away from traditional reliance on past experiences and enabling more data-driven agricultural practices [7].

A comprehensive review by Condran et al. (2022) on the role of ML in agriculture further emphasizes the potential of these technologies for supply chain optimization. The review identified key areas where ML can provide value, particularly when integrated with technologies like the Internet of Things (IoT) to improve decision-making and real-time data collection [8].

In the context of Bangladesh, Moon et al. (2023) synthesized the performance of different ML techniques, including Random Forest (RF), Ridge Regression, Naïve Bayes, and Category Boosting (CatBoost), in predicting wheat and rice yields. The study evaluated the models based on their ability to handle large, high-dimensional agricultural datasets. Random Forest, an ensemble method, provided robust predictions by aggregating multiple decision trees, while Ridge Regression addressed multicollinearity by regularizing linear models. Naïve Bayes applied probabilistic reasoning for classification, and CatBoost leveraged gradient boosting with categorical feature handling to enhance prediction accuracy. They suggested that future studies should incorporate larger, more diverse datasets to improve model accuracy, a recommendation highly applicable to the jute supply chain [9].

The multidimensional variables influencing jute production, such as climate changes, are not captured by traditional forecasting approaches. The majority of current research concentrates on univariate forecasting. Furthermore, the underutilization of ML methods exacerbates supply chain unpredictability and inefficiencies, especially contexts with low resources like Bangladesh.

One promising approach to overcoming the limitations of traditional forecasting methods is the use of ensemble learning models. Ensemble models, such as RF, Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost), combine multiple ML algorithms to improve accuracy and robustness. According to Tripathi & Biswas (2024), these models are particularly effective in handling multivariate datasets, where several factors interact to influence outcomes. For example, in agricultural forecasting, ensemble learning can combine data on weather conditions, soil health, market prices, and logistical constraints to provide a more holistic and accurate prediction of supply chain dynamics [10]. Studies have shown that ensemble learning methods can reduce forecasting errors by 15-20% compared to individual models, making them a powerful tool for optimizing agricultural supply chains [11].

In the context of the jute industry, the adoption of ensemble learning models could lead to improvements in supply chain efficiency. Through the incorporation of a range of factors such as climate data, market trends, and transportation logistics, these models can provide more reliable predictions for jute production and demand, helping stakeholders make better-informed decisions [12]. The optimization of the jute supply chain through more accurate forecasting would not only stabilize the market but also reduce waste, improve resource allocation, and enhance the overall resilience of the industry [13].

Therefore, this study aims to optimize the jute supply chain in Bangladesh through leveraging ensemble learning algorithms for accurate multivariate forecasting. The suggested framework integrates historical data related to jute and climate to improve decision-making through reliable forecasts. This study also plans to address the following research questions (**RQs**):

RQ1: How can ensemble learning models be employed to forecast jute yield rate and production constructively?

RQ2: What are the implications of the results on the jute industry stakeholders in Bangladesh?

The study endeavors to accomplish the following research objectives (**ROs**) by addressing the **RQs:**

RO1: To develop and evaluate ensemble learning models for forecasting jute yield rate and production, and assess their performance.

RO2: To offer findings to help stakeholders and policymakers in Bangladesh's jute sector make decisions, maximize resources, and develop sustainable supply chain management plans.

The novelty of this study lies in its utilization of ML to jute forecasting, where a range of features are utilized to forecast yield rate and production within a single framework. This strategy might support the jute industry's long-term viability, ensuring that it remains a vital component of Bangladesh's economy in the face of escalating environmental changes and global competitiveness.

The remainder of this study is structured as follows: Section 2 outlines the methodology. Section 3 presents the results and analysis. Section 4 discusses the significance of findings, while the study concludes with Section 5.

2. Methodology

The framework proposed for multivariate energy forecasting in Bangladesh is illustrated in **Fig.1**.

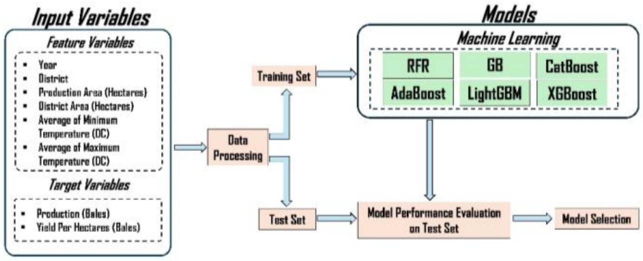

Fig.1: Framework for multivariate jute supply chain optimization in Bangladesh

The study comprises several steps: (1) data collection; (2) data processing, involving preprocessing to ensure data quality and feature engineering for meaningful insights; (3) partitioning the dataset; (4) utilizing the training set for model training; (5) assessing models on the test set; and (6) selecting the best model based on performance metrics.

2.1 Collection of data and analysis

Published by the Bangladesh Bureau of Statistics, historical public data are gathered from the Department of Agricultural Extension. Yearly district wise data as exhibited in **Table 2** are recorded as a Microsoft Excel file from 2019 to 2023.

Table 2 Accumulated data.

Feature No.	Data Collected	Data Type
1	Year	int64
2	District	object
3	Production Area (Hectares)	float64
4	Yield Per Hectares (Bales)	float64
5	Production (Bales)	int64
6	District Area (Hectare)	float64
7	Average of Minimum Temperature (DC)	float64
8	Average of Maximum Temperature (DC)	float64

Fig.2 is a graphical representation of the correlation matrix which helps to visualize correlation coefficients between each pair of features.

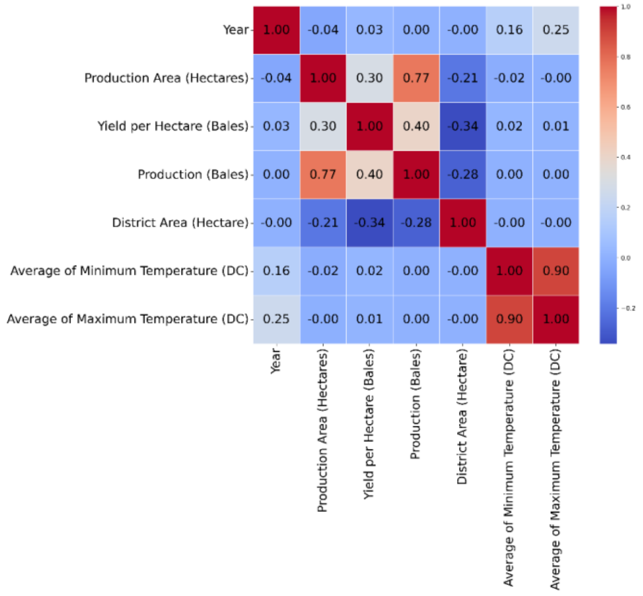

Fig.2: Correlation heatmap

Fig. (3 – 4) present line plots depicting trends in Yield Per Hectares (Bales) and Production (Bales) across the years 2019 to 2023. Fig.3 shows the yield per hectare (Bales) for various years from 2019 to 2023, with a general increasing trend and notable variability each year. Fig.4 shows the production (in Bales) for various years from 2019 to 2023, displaying high variability each year with a noticeable downward trend over the years.

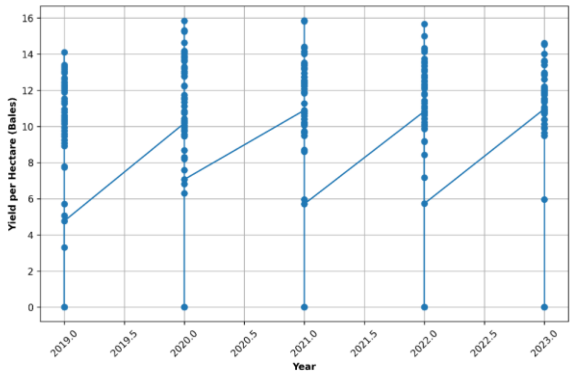

Fig.3: Year vs. yield per hectare (bales)

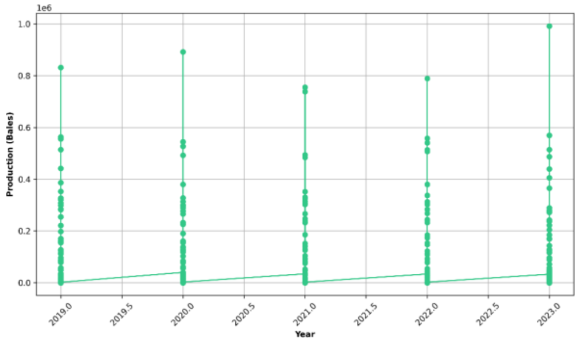

Fig.4: Year vs. production (bales)

2.2 Data processing

Given the time span, missing values are present within the dataset. Ensuring the quality and completeness of the experimental data is essential. Missing values are filled in using last observation carried forward (LOCF), as shown in **Eq. (1)**:

$$x_n = \begin{cases} x_n & \text{when } x_n \text{ is not missing} \\ x_{n-1} & \text{when } x_n \text{ is missing} \end{cases} \quad (1)$$

Let x_n represents the value at time m. When x_n is missing, it is substituted by x_{n-1}, the value from the preceding time step. LOCF propagates the last valid observation forward to fill the missing values along each column [14].

It is often common to split data 80:20 to have sets for use in training and testing where 80% of the data is used for training purposes while the rest of 20 % is used in testing. However, to maintain the chronological sequence, the 'shuffle' parameter is set to 'True.' The 'random_state' is also set to 42 to guarantee the reproducible nature of subsequent runs.

The features are normalized using 'StandardScaler' from 'scikit-learn' library, as demonstrated mathematically in Eq. (2):

$$X = \frac{M-N}{V} \quad (2)$$

where, transformed value of the feature is represented by X, the original value by M, the mean by N, and the standard deviation by V. Both the training and testing sets of features are scaled to maintain consistency between their scales. Similarly, the target variables are scaled independently to prevent any potential data leakage during model training.

2.3 Ensemble models

In this study, several ML models were utilized for forecasting, including Random Forest Regressor (RFR), Adaptive Boosting (AdaBoost), GB, Light Gradient Boosting (LightGBM), CatBoost, and XGBoost. RF operates by constructing multiple decision trees and averaging their outputs to enhance accuracy and reduce overfitting, making it robust against noise and effective for large datasets [15]. AdaBoost works by sequentially training weak classifiers and focusing on misclassified instances in each iteration, improving the performance of weak learners [16]. GB builds models in a sequential manner, with each model correcting the errors of the previous one, offering strong predictive performance, though it can be computationally intensive [17]. LightGBM, optimized for speed and efficiency, is particularly well-suited for large datasets, using a leaf-wise growth strategy to reduce training time while maintaining high accuracy [18]. CatBoost, designed to handle categorical data efficiently without extensive preprocessing, reduces overfitting through ordered boosting [19]. XGBoost, an optimized version of gradient boosting, is known for its high performance, scalability, and ability to handle sparse data, incorporating regularization to prevent overfitting [20]. These models were selected for their robustness in handling complex, non-linear relationships, providing high accuracy and efficiency in forecasting tasks.

2.4 Evaluation metrics for ML models

Five metrics have been utilized for comparison in order to evaluate the efficacy of suggested models. MSE, and coefficient of determination (R^2) indices are utilized to quantify the differences between the forecasted and actual values. The following **Eq. (3)** and **Eq. (4)** are the calculation formulas over n_{sample} samples:

$$MSE\ (y,\hat{y}) = \frac{1}{n_{sample}} \sum_{t=0}^{n_{sample}-1} (y_t - \hat{y}_t)^2 \quad (3)$$

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{t=0}^{n_{sample}-1} (y_t-\hat{y}_t)^2}{\sum_{t=0}^{n_{sample}-1} (y_t-y)^2} \quad (4)$$

$$where,\ \bar{y} = \frac{1}{n_{sample}} \sum_{t=0}^{n_{sample}-1} y_t$$

Here, y_t is the true value and $\hat{y}_t$ is the predicted value corresponding to the t-th sample.

3. Results and comparative analysis

Forecast accuracy indices for the suggested ML models are presented in **Table 3**.

Table 3 Indices of ML models.

Model	MSE	R^2
RFR	0.0942	0.9138
AdaBoost	0.1453	0.8695
GB	0.0974	0.9144
LightGBM	**0.0820**	**0.9199**
CatBoost	0.0947	0.9164
XGBoost	0.0934	0.9151

The distinct properties and learning techniques used by each algorithm are responsible for the variations in MSE and R^2 results among the ensemble models. LightGBM uses a leaf-wise growth method and had the lowest MSE (0.0820) and highest R2 (0.9199). This method effectively minimizes loss by concentrating on the most important splits, enabling the model to manage intricate relationships within the dataset. Strong performance was also shown by XGBoost and GB, which decrease bias by successively correcting errors through iterative optimization procedures. Their computation expenses, however, might make it more difficult for them to improve forecasts as effectively as LightGBM. On the other hand, CatBoost, which is intended for categorical data, works well with structured datasets but could have minor issues with data that is dominated by numbers. Higher MSE values were shown by models like RFR and AdaBoost, most likely as a result of their inherent properties. Forecasts from several decision trees are averaged by RF, which improves robustness but may lessen sensitivity to intricate, non-linear patterns. Similarly, in datasets that have intrinsic noise or variability, AdaBoost's emphasis on misclassified occurrences may result in overfitting.

3.1 Comparison of actual and forecasted values

Fig.5 compares the actual values and the forecasted values of the two target variables provided by the best model.

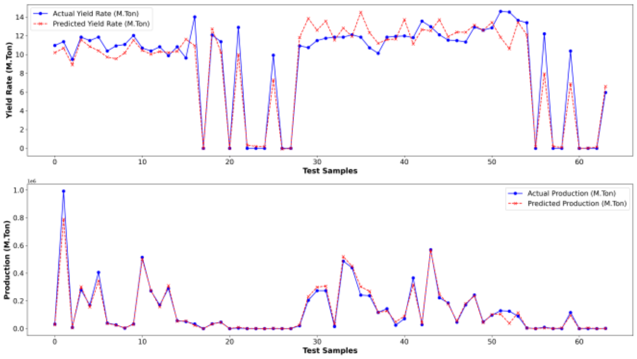

Fig.5: Comparison of actual and forecasted values

The upper plot contrasts the actual and forecasted yield rates for each test sample, which demonstrates the performance of LightGBM in identifying trends. The actual yield rates closely match the forecasted yield rates. The model's inability to adjust to sudden changes or noise in the dataset is probably the cause of the slight variations that are seen at extreme peaks and troughs. In spite of these difficulties, LightGBM handles yield rate forecasts with consistency.

Similarly, the lower plot contrasts the actual and forecasted production. The model's ability to capture intricate connections in the data while preserving consistency among test samples is demonstrated by the close alignment across different production values. This demonstrates the effectiveness with which LightGBM can generalize, even in situations with complex multivariate interactions.

4. Significance of the findings

The significance of this study lies in its contribution to optimizing the jute supply chain in Bangladesh through the application of advanced ML models, specifically LightGBM. With an impressive MSE of 0.0820 and an R^2 value of 0.9199, LightGBM has demonstrated exceptional accuracy in predicting jute production and demand, outperforming traditional forecasting methods. This study is particularly important for addressing the inefficiencies and volatility that have long plagued the jute industry, from unpredictable weather patterns to fluctuating market demands. Through providing a more accurate forecasting model, this research offers the potential to stabilize the jute supply chain, reduce wastage, and improve resource allocation. This, in turn, could lead to better decision-making among farmers, traders, and policymakers, ensuring higher profitability and sustainability in the industry. Furthermore, the study's focus on multivariate forecasting using ensemble learning can serve as a model for other agricultural sectors in Bangladesh, paving the way for more data-driven approaches to crop management and supply chain optimization.

5. Conclusion

The study demonstrates the potential of multivariate ensemble learning algorithms to optimize the jute supply chain in Bangladesh through highly accurate forecasts for yield rates and production. The results of LightGBM highlight the benefits of the integration of multivariate datasets and potential of ensemble learning models to enhance forecasting accuracy and reduce inefficiencies in the supply chain. Through leveraging these advanced ML techniques, stakeholders in the jute industry can make more informed decisions, leading to improved supply chain efficiency, reduced wastage, and increased profitability.

5.1 Theoretical implications

1. The ability of ensemble learning models to handle multivariate data with high accuracy is demonstrated in this study, which promotes their use in agricultural forecasting.
2. It adds to the body of knowledge on multivariate forecasting in agriculture by establishing a systematic framework for forecasting production and yield rate.
3. The study expands the theoretical understanding of tailored ML applications in emerging economies by

demonstrating the flexibility of LightGBM and related models in resource-constrained environments.

5.2 Practical implications

1. The increased forecasting accuracy can be used to minimize waste, maximize resource allocation, and make well-informed management decisions.
2. The model's forecasts can be used by supply chain stakeholders to lessen market volatility and supply chain instability.
3. Adoption of data-driven supply chain optimization techniques in agriculture is encouraged by the framework's possible implementations for different crops and geographical areas.

5.3 Limitations and future work

Despite preprocessing efforts, this study is limited by its reliance on historical data, which may contain inherent discrepancies. The results are specific to the jute sector, which restricts their applicability to other crops or areas without changes. Algorithms' computing requirements may also make large-scale adoption difficult in environments with limited resources.

To further enhance the results, future studies could broaden datasets to incorporate more contextual and temporal factors, like foreign market trends and transportation costs. Incorporating predictive models into decision-support systems and conducting comparative studies across various crops and geographical areas may ensure wider scalability and practical utilization.

References:

[1] Roy, S., & Lutfar, L. B. (2012). Bast fibres: jute. In Handbook of natural fibres (pp. 39-59). Woodhead Publishing.

[2] Hossain, M. M., & Abdulla, F. (2015). Jute production in Bangladesh: a time series analysis. Journal of Mathematics and Statistics, 11(3), 93-98.

[3] Rahman, M., & Khaled, N. (2011). Global market opportunities in export of jute (No. 93). Centre for Policy Dialogue (CPD).

[4] Moazzem, K. G., Rahman, M. T., & Sobhan, A. (2009). Jute manufacturing sector of Bangladesh: challenges, opportunities, and policy options. Dhaka: Centre for Policy Dialogue.

[5] Tariq, M., Khan, M. A., Muhammad, W., & Ahmad, S. (2023). Fiber crops in changing climate. In Global Agricultural Production: Resilience to Climate Change (pp. 267-282). Cham: Springer International Publishing.

[6] Dahikar, S. S., & Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), 683-686.

[7] Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017, July). Agricultural production output prediction using supervised machine learning techniques. In 2017 1st international conference on next generation computing applications (NextComp) (pp. 182-187). IEEE.

[8] Condran, S., Bewong, M., Islam, M. Z., Maphosa, L., & Zheng, L. (2022). Machine learning in precision agriculture: a survey on trends, applications and evaluations over two decades. IEEE Access, 10, 73786-73803.

[9] Moon, S., Lee, S., Jeon, W., & Park, K. J. (2023). Learning-Enabled Network-Control Co-Design for Energy-Efficient Industrial Internet of Things. IEEE Transactions on Network and Service Management.

[10] Tripathi, D., & Biswas, S. K. (2024). Design of a precise ensemble expert system for crop yield prediction using machine learning analytics. Journal of Forecasting.

[11] Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18-41.

[12] Nti, I. K., Zaman, A., Nyarko-Boateng, O., Adekoya, A. F., & Keyeremeh, F. (2023). A predictive analytics model for crop suitability and productivity with tree-based ensemble learning. Decision Analytics Journal, 8, 100311.

[13] Mohamed-Iliasse, M., Loubna, B., & Abdelaziz, B. (2020, October). Is machine learning revolutionizing supply chain?. In 2020 5th International Conference on Logistics Operations Management (GOL) (pp. 1-10). IEEE.

[14] Hadeed, S. J., O'rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, *730*, 139140.

[15] Kulkarni, V. Y., Sinha, P. K., & Petare, M. C. (2016). Weighted hybrid decision tree model for random forest classifier. Journal of The Institution of Engineers (India): Series B, 97, 209-217.

[16] Joly, A. (2017). Exploiting random projections and sparsity with random forests and gradient boosting methods--Application to multi-label and multi-output learning, random forest model compression and leveraging input sparsity. arXiv preprint arXiv:1704.08067.

[17] Andreas, Mayr., Harald, Binder., Olaf, Gefeller., Matthias, Schmid. (2014). 1. The evolution of boosting algorithms. From machine learning to statistical modelling. Methods of Information in Medicine, doi: 10.3414/ME13-01-0122

[18] Shi, Y., Li, J., & Li, Z. (2018). Gradient boosting with piece-wise linear regression trees. *arXiv preprint arXiv:1802.05640*.

[19] Sprangers, O., Schelter, S., & de Rijke, M. (2021, August). Probabilistic gradient boosting machines

for large-scale probabilistic regression. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 1510-1520).

[20] Just, A. C., Liu, Y., Sorek-Hamer, M., Rush, J., Dorman, M., Chatfield, R., ... & Kloog, I. (2020). Gradient boosting machine learning to improve satellite-derived column water vapor measurement error. Atmospheric measurement techniques, 13(9), 4669-4681.

NOMENCLATURE

X^n	: n-dimensional input space
n_{sample}	: Total number of samples
y_t	: True value corresponding to the t-th sample
$\hat{y}_t$	: The predicted value corresponding to the t-th sample
X	: The standardized value
M	: The original value
N	: The mean of the feature
V	: The standard deviation