Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

Enhancing Agricultural Price Forecasting using Ensemble Models and Optimized Feature Selection Techniques

D. Lawanya[1], Dr. N. Muthumani[2]

[1]Research scholar, Department of computer science, Ppg college of arts and science, lawanyaagriml@gmail.com

[2]Principal, Department of computer science Ppg college of arts and science, saravanampatty, Coimbatore

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract: Agricultural commodity price fluctuations have garnered significant attention due to their economic impact. This research addresses the challenges of predicting agricultural commodity prices by leveraging advanced machine learning techniques. Unlike traditional statistical models, ensemble-based machine learning algorithms are employed to better capture the nonlinear and complex dynamics of price series. To enhance predictive accuracy, this study incorporates feature selection methods, including Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Genetic Algorithm (GA), and Grey Wolf Optimization (GWO), for dimensionality reduction and relevant feature extraction. The selected features are utilized by four ensemble classifiers: AdaBoost, XGBoost, CatBoost, and Gradient Boosting. The proposed model is evaluated using key performance metrics such as Precision, Recall, Kappa, Root Mean Square Error (RMSE), and Accuracy. Experimental results demonstrate that the proposed methodology significantly improves forecasting performance, providing a robust alternative to traditional approaches.

Keywords: Machine learning, Ensemble learning, Agriculture commodity, Classification, Price forecasting.

1. Introduction

Agricultural commodities play a vital role in both local and global economies, influencing food security and international trade [1-3]. Accurate price prediction is essential for stabilizing farmer incomes, mitigating market risks, and supporting informed policy decisions to manage food supply chains and reduce market volatility [4, 5]. However, agricultural prices are inherently volatile, driven by unpredictable factors such as climate conditions, seasonal variations, and macroeconomic trends, making forecasting a significant challenge [6].

Traditional forecasting methods, such as econometric models and regression techniques, have struggled to address these complexities. These approaches often fail to process high-dimensional, nonlinear data and capture the intricate patterns within the agricultural sector [7-9]. In response, machine learning techniques, particularly ensemble learning, have gained traction due to their ability to enhance prediction accuracy and generalization by leveraging multiple algorithms [10, 11].

This study proposes an advanced approach that integrates feature selection techniques with ensemble learning to improve agricultural commodity price predictions. Feature selection methods—including Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Genetic Algorithm

(GA), and Grey Wolf Optimization (GWO)—are utilized to extract the most relevant features, reducing noise and computational complexity while enhancing model performance [12-17].

To build a robust predictive framework, four ensemble-based classifiers—AdaBoost, XGBoost, CatBoost, and Gradient Boosting—are employed. These classifiers, known for their ability to transform weak learners into strong predictors, are evaluated using key performance metrics such as precision, recall, accuracy, and Root Mean Square Error (RMSE) [18-20]. By integrating advanced feature selection with ensemble learning, this study presents a novel, high-performance methodology for agricultural price forecasting. Extensive experimentation and comparative analysis validate the effectiveness of this approach, offering valuable insights into enhancing prediction accuracy and developing reliable forecasting tools for real-world agricultural applications.

2. Literature Review

Predicting agricultural commodity prices remains a significant challenge due to the complex and dynamic nature of the agricultural sector. Traditional forecasting techniques, such as econometric models, time-series analysis, and statistical methods, have been widely used. However, these approaches often struggle to account for the non-linearity and high dimensionality inherent in agricultural datasets. With advancements in machine learning (ML), researchers have increasingly explored ML-based methods to enhance the accuracy and robustness of commodity price predictions.

Abdullah [21] proposed a hybrid model for forecasting coconut prices, addressing the challenges posed by price fluctuations. The model integrates ARIMA and ANN techniques to leverage both linear and nonlinear modeling capabilities. An experimental study demonstrated that combining ARIMA and ANN improves forecasting accuracy and aids in identifying trends in coconut price data.

Mohanty et al. [22] introduced a machine learning-based framework for crop price prediction, helping farmers estimate profit and loss in advance. The framework consists of four key components: crop yield prediction, supply-demand estimation, and price forecasting. Various methods, including machine learning, statistical approaches, and time series models, were applied to predict prices based on supply, demand, and time trends. A comparative analysis identified the decision tree regressor as the most effective model, achieving the lowest root-mean-square error (RMSE).

Avinash et al. [23] developed a Hidden Markov Model (HMM)-guided deep learning approach for forecasting nonlinear and nonstationary agricultural commodity price data. By incorporating technical indicators, this approach enhances forecasting precision, benefiting stakeholders such as farmers and policymakers.

Zhang and Tang [24] proposed a novel VMD-SGMD-LSTM model that integrates artificial intelligence with advanced quadratic decomposition techniques. Initially, the futures price data undergoes decomposition using Variational Mode Decomposition (VMD) and further refinement through SGMD. The final predictions are generated by aggregating the expected values from different modal components, predicted using an LSTM model.

Rana et al. [25] explored the application of a big data framework for agricultural price forecasting in Pakistan. Using a historical dataset (2007–2022) on commodity prices across various cities and

employing Apache Spark for data preprocessing, the study applied ARIMA, Random Forest, and Long Short-Term Memory (LSTM) models to forecast price trends. Results indicated that LSTM outperformed ARIMA and Random Forest, achieving an R^2 value of 0.8 and the lowest Mean Absolute Error (MAE) of 125.29, demonstrating its superior predictive capabilities.

Sun [26] investigated garlic price forecasting in Jinxiang, China. The study first extracted features using VMD decomposition, generating a combined feature set (De_Vo) by incorporating volatility indicators. Classification models, including logistic regression, SVM, and XGBoost, were employed to predict price trends. Results showed that feature-enhanced models outperformed individual feature-based predictions, with XGBoost achieving the highest accuracy (72.9%), followed by SVM (71.4%) and logistic regression (62.6%).

The paper is structured as follows: Section 2 details the methodology and framework design. Section 3 presents the results, analysis, discussion, and evaluation methods. Finally, Section 4 summarizes the study's conclusions and outlines future research directions.

3. Methods

The proposed approach demonstrates strong potential for improving the accuracy of agricultural commodity price forecasting by integrating various feature selection techniques with machine learning models. This methodology has significant implications for stakeholders such as farmers and policymakers, aiding in better decision-making and market planning. Initially, the collected data undergoes preprocessing using label encoding. Subsequently, different feature selection algorithms—including Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and Genetic Algorithm (GA)—are applied to identify the most relevant features, thereby enhancing classifier performance. Finally, the impact of these three feature selection techniques on ensemble machine learning classifiers is analyzed and evaluated. The overall architecture of the proposed agricultural commodity price prediction model is illustrated in Figure 1.

3.1 Data Collection

The dataset used in this research consists of historical price data for various agricultural commodities, including crop prices such as wheat, rice, and corn. The dataset was sourced from publicly available agricultural databases, including [specify data sources], which provide daily, weekly, or monthly price data along with associated features such as weather conditions, market demand, supply chain data, and macroeconomic indicators.

3.2 Preprocessing

The dataset underwent various operations related to data cleaning and preparation, including column renaming, duplicate and superfluous column removal, handling missing values, forward filling null values with the previous non-null value, and data type conversion.

Missing values were imputed using median imputation for numerical variables [27] and mode imputation for categorical variables. Min-Max scaling [28] was applied to normalize the dataset, ensuring that all features had values between 0 and 1. This prevents bias due to varying feature scales. Categorical variables were converted into numerical values using one-hot encoding.

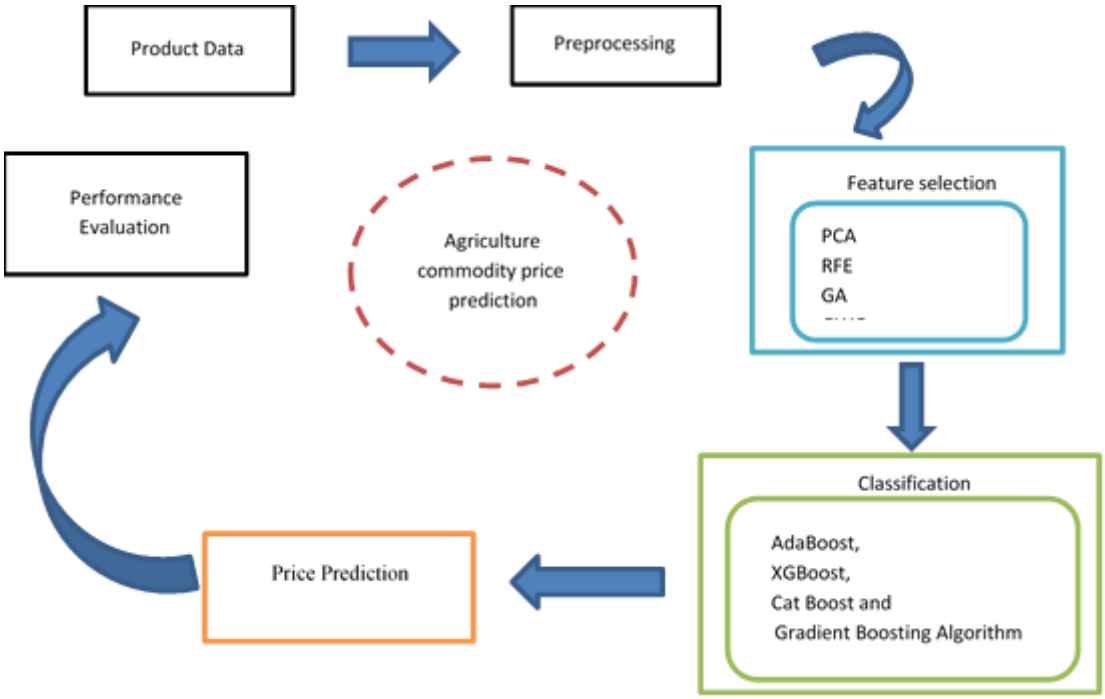

Figure 1. Agriculture commodity price prediction model

3.3 Feature selection:

3.3.1 Principle component analysis

PCA is a linear dimensionality reduction method that projects data into a lower-dimensional sub-space, allowing for the extraction of information from a high-dimensional environment. It attempts to eliminate the non-essential sections with less variation and keep the vital parts with more variation in the data. The algorithm steps of PCA is given in table 1.

Being an unsupervised dimensionality reduction technique, PCA can cluster comparable data points based on the feature correlation between them without the need for supervision or labeling. This is an essential observation regarding PCA. Coordinate rotation and transformation constitute the mathematical core of the PCA approach. In the new coordinate system, the original n variables are linearly integrated to create n new variables that are unrelated to one another. The essential steps of the PCA approach are stated as follows, using the slope with only two nodes in the finite element simulations as an example for simplicity.

Table 1: Algorithm of PCA

Algorithm: PCA
Step 1: Centralize the cohesion matrix
Step 2: compute the covariance matrix
Step 3: Determine the covariance matrix's eigenvalues and eigenvectors.
Step 4: Choose a benchmark for dimensionality reduction
Step 5: Calculate the dataset after dimensionality reduction

3.3.2 Recursive Feature Elimination

One feature selection technique for determining the important characteristics in a dataset is recursive feature elimination. Once the least important components have been eliminated several times, the procedure entails creating a model with the remaining characteristics until the required number of features is reached. RFE can be applied to any supervised learning technique [29]. The steps involved in the RFE is given in table 2.

The RFE was a wrapper type. This indicates that a distinct machine learning algorithm is provided, utilized as the central component of the method, wrapped with RFE, and used to aid in feature selection. On the other hand, filter-based feature selections assign a score to every feature and then choose the features that have the highest (or lowest) value. When utilizing RFE, two key configuration options are considered: the number of features to choose, and the algorithm to assist in feature selection. RFE discovers a subset of features by first utilizing every feature in the training dataset and then effectively eliminating the features until the target number of features is retained.

Table 2: Algorithm of RFE

Algorithm: RFE
Step 1: Using the selected RFE machine learning algorithm, rank the significance of each feature.
Step 2: Remove the least significant element.
Step 3: build a model using the remaining features.
Step 4: Apply the dataset on the model
Step 5: Until the required number of features is attained, repeat steps 1-3.

3.3.3 Genetic Algorithm

The goal of GA is to assess the psychological effects, model the variable approaches, and mimic the natural changes that take place in social systems, which are living ecosystems [30]. GA provides a sizable number of issues may essentially be resolved by applying the GA techniques. GA is a well-liked search and optimization technique for handling extremely complex issues. Its techniques have shown to be successful in fields where machine learning is used. This section contains a detailed description of the actual coded GA. The flow chart of the GA is illustrated in figure 2.

The conventional GA process is described below.

Initial population

This entails the possible solution for set G, i.e., a series of random generations of real values,

$$G = \{g1, g2, \dots, gs\}.$$

Evaluation

To assess every chromosome in the population, the fitness function, which is defined as $fitness = g(P)$, must be defined.

Selection

The chromosomes are sorted according to their fitness values after the fitness value computation. The next step is to pick the parents, which requires two parents for the crossover and the mutation.

The genetic operators are used to construct the children (C1, C2) or the parents' new chromosomes when the selection procedure is finished. After then, children in population C are spared with the new chromosomes (C1, C2). The crossover and mutation operations are involved in this process. Two parents that were chosen earlier exchange information using the crossover process. There are numerous crossover operator techniques, including arithmetical, two-, k-, and single-point crossovers, among others. The crossing offspring's chromosomes' genes are altered during the mutation procedure. Similarly, the mutation operator has multiple options.

Children population C is fully formed and will be transferred to the next population (P) after the selection, crossover, and mutation operations are finished. The method is then repeated using P in the subsequent iteration. If the number of iterations exceeds the maximum threshold or if the results start to converge, the iterations will end.

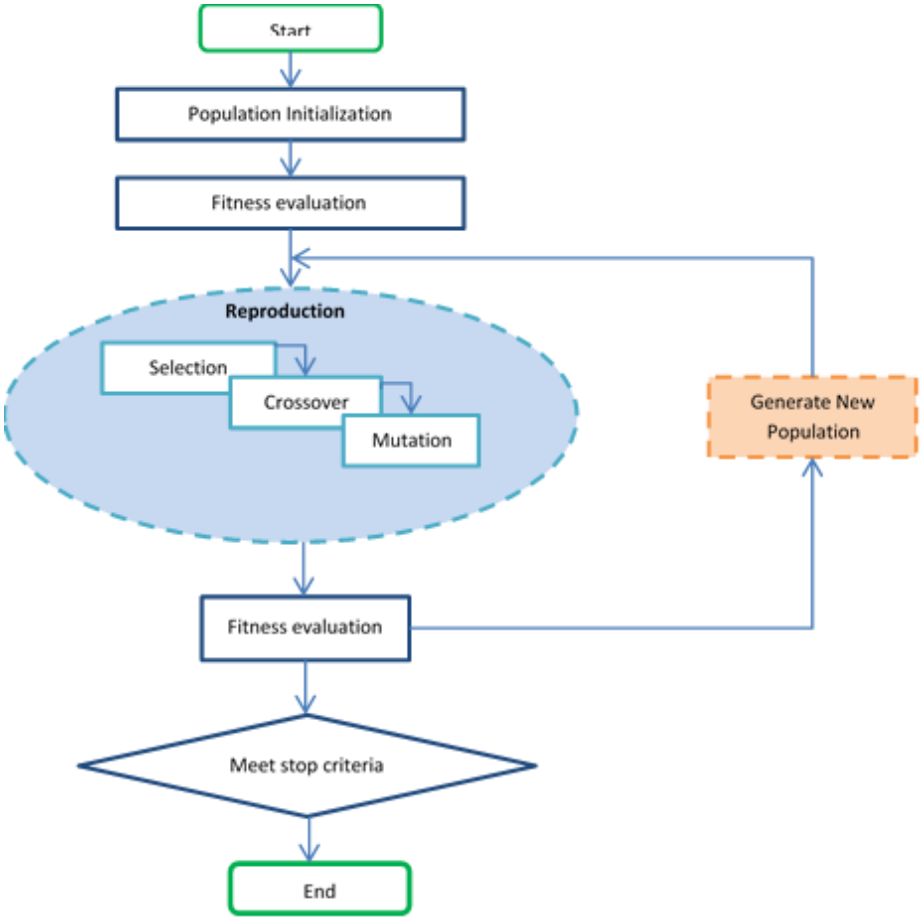

Figure 2. Flow chart of Genetic Algorithm

3.3.4 Grey wolf optimization

Grey Wolf Optimization (GWO) is a nature-inspired metaheuristic algorithm that mimics the leadership hierarchy and hunting strategy of grey wolves in the wild. It was introduced by Seyedali

Mirjalili in 2014 [31] and is widely used for solving optimization problems. The algorithm is inspired by the social structure of grey wolves, which typically hunt in packs, led by four types of wolves: alpha (leader), beta (second in command), delta, and omega (followers). GWO models this hierarchy and utilizes it to find optimal solutions by simulating the wolves' cooperative behavior in hunting prey. Pseudocode of GWO is given in table 3.

Initialization:

Initialize the population of grey wolves (candidate solutions). Each wolf represents a potential solution to the optimization problem. Identify the alpha (best solution), beta (second-best solution), and delta (third-best solution) wolves based on their fitness values (objective function).

Update Position of Grey Wolves:

For each wolf (solution) in the population, update its position based on the influence of alpha, beta, and delta wolves.The position update formula is defined by

$$X(a+1)\frac{X_\alpha(a)+X_\beta(a)+X_\gamma(a)}{3}$$

Where $X(a)$) is the current position of the wolf. X_α,X_β,X_γ are the positions of the alpha, beta, and delta wolves, respectively. The position is updated based on the wolves' distances from the best solutions.

Encircling Prey:

The grey wolves attempt to encircle their prey (optimal solution) by adjusting their positions in relation to alpha, beta, and delta wolves. The encircling behavior is mathematically modeled as

$$R = |Q.X_b(a) - X(a)|$$

$$X(a+1) = X_b(a) - P.Q$$

Where R is the distance between the wolf and the prey, P and Q are coefficient vectors that dynamically adjust the influence of alpha, beta, and delta wolves on the position update and $X_b(a)$ denotes the position of the best solution found so far.

Coefficient Updates:

The values of P and Q are updated during the optimization process to balance exploration and exploitation:

$$P = 2t.r_1 - t$$

$$Q = 2.r_2$$

Where t denotes the decreases linearly from 2 to 0 during the iterations, controlling the exploration-exploitation trade-off, r_1 and r_2 are random vectors in [0, 1] to introduce stochastic behavior.

Exploration and Exploitation:

Exploration: When $|P| > 1$, the wolves search the solution space more broadly, encouraging exploration.

Exploitation: When $|P| < 1$, the wolves focus on refining their positions around the best solutions found, enabling exploitation.

Fitness Evaluation:

Evaluate the fitness (objective function value) of each grey wolf based on the current position (solution). Update the positions of the alpha, beta, and delta wolves if new better solutions are found.

Termination:

Repeat the position updating, encircling, and fitness evaluation steps until the maximum number of iterations is reached or the stopping criteria (e.g., convergence) are satisfied. The alpha wolf (best solution) at the end of the optimization process is considered the optimal solution.

Table 3: Pseudocode of GWO:

Pseudocode of GWO:

Initialize population of grey wolves (solutions)
Initialize maximum number of iterations (MaxIter)
Initialize the alpha, beta, and delta wolves
For each iteration (t = 1 to MaxIter):
  For each grey wolf:
  Update position using alpha, beta, and delta positions
  Encircle the prey by updating position vectors (A, C)
  Update coefficients A and C
  Evaluate the fitness of each grey wolf
  Update alpha, beta, and delta based on fitness values
  If stopping condition is met, break
Return the alpha wolf as the best solution

3.4 Classification:

3.4.1 AdaBoost

An AdaBoost classifier starts by fitting a copy of the original dataset using a copy of the same classifier that has been updated to remove error-prone and inaccurate data points. This allows the subsequent classifiers to concentrate on the cases that lead to greater inaccuracy [32].

AdaBoost is a type of iterative calculation whose basic idea is to prepare multiple classifiers (that is, weak classifiers) using a preparation set, and then use several different strategies to coordinate them to create a more grounded classifier. The computation itself is carried out by adjusting the information flow, as demonstrated by the preparation set test's order amendment and the final accuracy of the overall arrangement to determine each example's weight. Subsequently, forward the

updated data to the lower classifier for preparation. Ultimately, each preparation classifier is combined to generate the official conclusion classifier.

1. The transmission of the example, rather than the resampling, is the focal point of each modification;
2. The misclassified test weight high and the arranged effectively test weight low determine the difference in test dispersion. This will enable the subsequent classifier to be centered around the current misclassified tests;
3. To obtain the result, add together all of the weak classifiers [33].

As a result, it is seen that AdaBoost concentrates on the incorrectly identified points by giving the misclassified data greater weights, which lowers error and raises accuracy.

3.4.2 XGBoost

Extreme Gradient Boost is referred to as XGBoost. The boosting calculation's basic idea is that several decision trees outperform a single one. Not every decision tree will present well. The presentation starts to get better at the point where several trees are added.

The steps involved in XGBoost:

1. Construct a standard set that is strongly connected, and if any of the elements in the prepared set are constant, discretize the component vectors.
2. After determining which principles produce the forecast mark, create the new standard set. Determine the lift of every standard and remove the principles if the lift is less than 1.0. We can obtain the strong set of guidelines in this manner.
3. The concepts in the successful standard are arranged according to their greatest assistance, lift, and brief length.
4. Print the item from the capabilities list and exit if the cycle record is larger than the specified amount. If not, we extract the principal rule from the effective standard set, append its outcome to the capabilities list, and remove the standard from the set of workable principles. The component that is now listed among the capabilities doesn't need to be included.
5. Remove instances that meet the requirements, then determine lift and backing for the remaining preparation set.
6. Go to Step 4 after sorting the requirements according to the workable principle of maximum assistance, most extreme lift.

XGBoost is an inclination boosting system-based decision tree-based machine learning calculation. In forecast problems involving unstructured data (images, text, etc.), artificial neural networks will typically outperform all other computations and structures. Nevertheless, decision tree based computations are thought to be the most effective for little to medium organized/forbidden data.

3.4.3 Gradient Boost

Gradient Boosting is a machine learning technique used for regression and classification problems, which builds a predictive model in a stage-wise fashion. It creates a strong predictive model by combining multiple weak learners, typically decision trees. The main idea is to correct the errors made by the previous models, effectively boosting their performance. The algorithm works by

sequentially adding models, where each new model is trained to predict the residuals (errors) of the existing ensemble of models. The final prediction is obtained by aggregating the predictions of all models.

3.4.4 CatBoost

CatBoost is a gradient-boosting decision tree (GBDT) architecture that uses fewer parameters and an oblivious tree as the basis learner. It attains great accuracy while supporting categorical variables. Uses the boosting approaches to train a sequence of learners serially, accumulating the outputs of all learners as a result [34-35], increasing the algorithm's accuracy and applicability. Given a training set containing n samples, D$\{(A_x, B_x)_{x=1,2,\cdots,n}\}$

where $B_x = (b_x^1, b_x^2, \cdots, b_x^n)$ stands for labeled values and $B_x \in R$ stands for the m-dimensional input features. The next training round's objective is to select a tree t_r from the CART decision tree set T in order to minimize the expectation E(·) of the loss function L(·), with the strong learner created after training being V_{r-1}. Here is how the parameter t_r is computed

Where test samples (a, b) is not part of the training set. The trained CART decision tree t_r is fitted by the GBDT using the negative gradient of the loss function and after N iterations, the final model M, represented in Equation (2), is produced from the initial weak learner W_0 and the n-th round of the training step size

4. Experimental Result

This section, evaluate and discuss the performance of the ensemble learning models (AdaBoost, XGBoost, CatBoost, and Gradient Boosting) applied to agriculture commodity price prediction, using multiple feature selection techniques. The dataset was split into two subsets: 80% for training and 20% for testing. A 5-fold cross-validation technique was employed on the training data to ensure model robustness and to avoid overfitting. The models were evaluated using several performance metrics, including Accuracy, Precision, Recall, F1 Score, Root Mean Squared Error (RMSE), Cohen's Kappa, and Matthews Correlation Coefficient (MCC).

Table 4. Performance Metrics of Ensemble Classifiers with different feature selection methods

Feature Selection Method	Classifier	Precision	Recall	F1 score	Accuracy	RMSE	MCC	Kappa
Principal Component Analysis (PCA)	AdaBoost	82.02	81.95	81.93	81.94	0.42	0.63	0.63
	Gradient Boosting	83.48	83.35	83.31	83.33	0.40	0.66	0.66
	XGBoost	76.49	74.64	74.26	74.72	0.50	0.51	0.49
	CatBoost	84.25	84.18	84.15	84.16	0.39	0.68	0.68
Recursive Feature	AdaBoost	81.10	82.07	81.10	81.11	0.43	0.62	0.62

Elimination (RFE)								
	Gradient Boosting	75.67	76.66	76.66	76.66	0.48	0.53	0.53
	XGBoost	75.13	74.24	73.80	74.0	0.50	0.49	0.48
	CatBoost	82.01	82.00	82.00	82.0	0.42	0.63	0.63
Genetic Algorithm (GA)	AdaBoost	81.10	81.07	81.10	81.11	0.43	0.62	0.62
	Gradient Boosting	76.67	76.66	76.66	76.66	0.48	0.53	0.53
	XGBoost	75.13	74.24	73.80	74.0	0.50	0.49	0.48
	CatBoost	82.01	81.00	82.00	82.0	0.42	0.63	0.63
Grey Wolf Optimization (GWO)	AdaBoost	87.80	87.79	87.77	87.77	0.34	0.75	0.75
	Gradient Boosting	85.66	85.59	85.55	85.55	0.38	0.71	0.71
	XGBoost	80.10	77.95	77.73	78.14	0.46	0.58	0.56
	CatBoost	88.15	88.15	88.14	88.14	0.34	0.76	0.76

Table 4 provides a comparison of the performance of different feature selection methods combined with various classifiers across several evaluation metrics. And the figure 3- 6 shows the graphical visualization of the result achieved. Each combination of feature selection method and classifier is evaluated based on metrics like precision, recall, F1 score, accuracy, RMSE, MCC, and Kappa, which assess the quality and performance of the models.

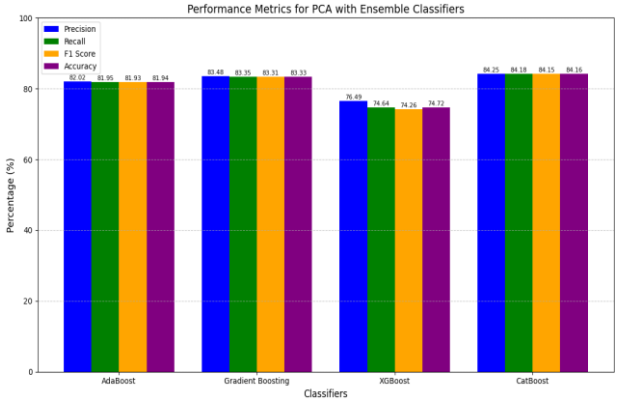

Figure 3. Ensemble model performance with PCA feature selection

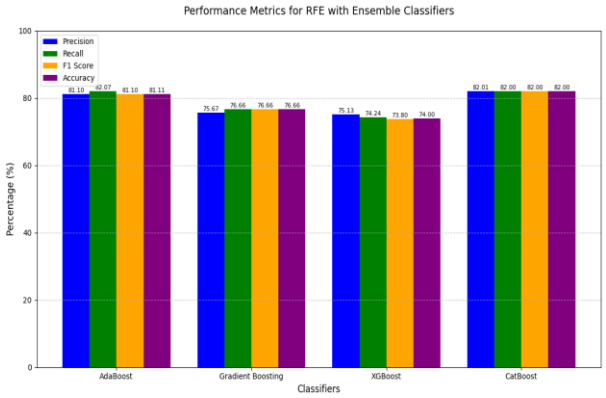

Figure 4 Ensemble model performance with RFE feature selection

4.1 Principal Component Analysis (PCA) + Classifiers

The evaluation of PCA with different ensemble classifiers shown in figure 3 reveals that CatBoost outperforms the others, achieving the highest precision (84.25%), recall (84.18%), F1 score (84.15%), and accuracy (84.16%) with the lowest RMSE (0.39) and robust MCC and Kappa scores (0.68). Gradient Boosting follows closely with slightly lower metrics but still demonstrates strong reliability, while AdaBoost provides moderate performance, achieving an accuracy of 81.94% and MCC/Kappa scores of 0.63. XGBoost, however, underperforms with the lowest accuracy (74.72%), highest RMSE (0.50), and weaker MCC/Kappa scores, indicating its limited compatibility with PCA-selected features. This analysis highlights CatBoost as the most effective classifier for agricultural commodity price prediction when paired with PCA, showcasing its ability to generalize well and handle categorical features efficiently.

4.2 Recursive Feature Elimination (RFE) + Classifiers: The performance evaluation of Recursive Feature Elimination (RFE) with various ensemble classifiers shown in figure 4 highlights CatBoost as the best-performing model, achieving precision, recall, and F1 score of 82.0%, along with an accuracy of 82.0%, the lowest RMSE (0.42), and reliable MCC and Kappa scores (0.63). AdaBoost follows closely, with slightly lower metrics, including an accuracy of 81.11% and MCC/Kappa scores of 0.62, showing moderate reliability. Gradient Boosting achieves an accuracy of 76.66% but lags behind with a higher RMSE (0.48) and lower MCC/Kappa scores (0.53). XGBoost underperforms, with the lowest accuracy (74.0%), the highest RMSE (0.50), and weak MCC/Kappa scores, indicating limited compatibility with RFE. Overall, this analysis emphasizes CatBoost's robustness when combined with RFE for agricultural commodity price prediction, while AdaBoost also proves to be a reliable alternative.

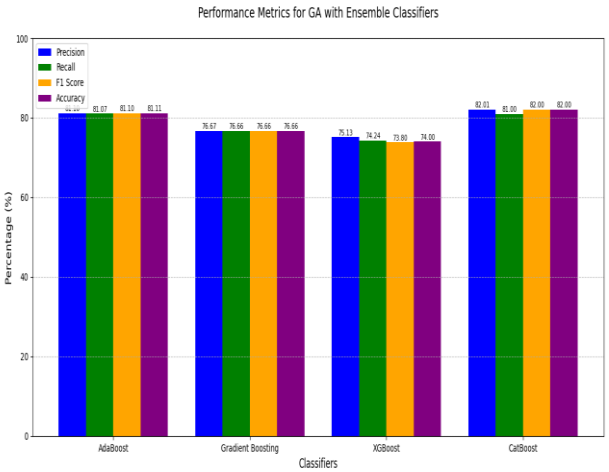

Figure 5. Ensemble model performance with GA feature selection

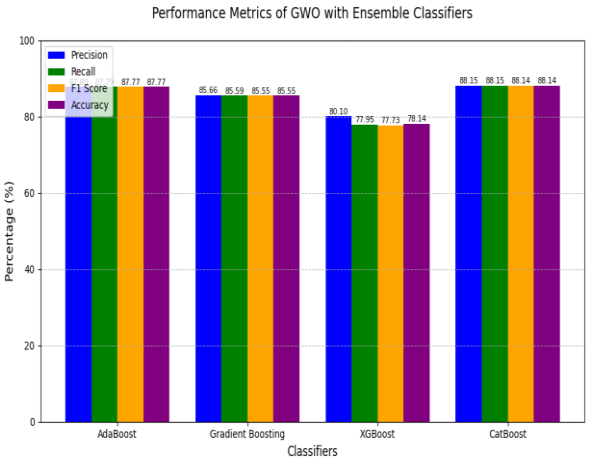

Figure 6 Ensemble model performance with GWO feature selection

4.3 Genetic Algorithm (GA) + Classifiers:

The evaluation of Genetic Algorithm (GA) with ensemble classifiers shown in figure 5 highlights CatBoost as the top performer, achieving precision, recall, and F1 score of 82.0%, accuracy of 82.0%, the lowest RMSE (0.42), and strong MCC/Kappa scores (0.63). AdaBoost closely follows with an accuracy of 81.11%, a slightly higher RMSE (0.43), and MCC/Kappa scores of 0.62,

demonstrating moderate reliability. Gradient Boosting performs reasonably with an accuracy of 76.66%, though its higher RMSE (0.48) and lower MCC/Kappa scores (0.53) suggest room for improvement. XGBoost again underperforms, showing the lowest accuracy (74.0%), the highest RMSE (0.50), and weak MCC/Kappa scores, indicating limited effectiveness when combined with GA. This analysis confirms CatBoost's robustness with GA for agricultural commodity price prediction, while AdaBoost remains a reliable alternative.

4.4 Grey Wolf Optimization (GWO) + Classifiers:

The performance evaluation of Grey Wolf Optimization (GWO) combined with ensemble classifiers shown in figure 6 highlights CatBoost as the best performer, achieving the highest precision (88.15%), recall (88.15%), F1 score (88.14%), and accuracy (88.14%), along with the lowest RMSE (0.34) and the strongest MCC and Kappa scores (0.76). AdaBoost follows closely, with slightly lower metrics, including an accuracy of 87.77% and MCC/Kappa scores of 0.75, showcasing strong reliability. Gradient Boosting achieves moderate performance, with an accuracy of 85.55% and MCC/Kappa scores of 0.71, indicating good but less robust results compared to CatBoost and AdaBoost. XGBoost, however, lags behind with the lowest accuracy (78.14%), the highest RMSE (0.46), and weaker MCC/Kappa scores (0.58), reflecting limited compatibility with GWO. This analysis emphasizes CatBoost's superior performance and reliability with GWO for agricultural commodity price prediction, with AdaBoost as a close alternative.

5. Discussion

This study evaluates the effectiveness of various feature selection techniques—Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Genetic Algorithm (GA), and Grey Wolf Optimization (GWO)—in combination with ensemble classifiers, including AdaBoost, Gradient Boosting, XGBoost, and CatBoost, for agricultural commodity price prediction. The objective is to determine the optimal combination of feature selection and classification methods to achieve the highest predictive accuracy.

Key findings based on performance metrics such as precision, recall, F1 score, accuracy, RMSE, MCC, and Kappa are as follows:

Best Feature Selection Method: Grey Wolf Optimization (GWO) consistently outperforms other techniques, achieving the highest precision, recall, F1 score, and accuracy across all classifiers. This demonstrates its effectiveness in selecting relevant features and minimizing noise, making it the most suitable feature selection method for agricultural price prediction.

Best Classifier: CatBoost emerges as the top-performing classifier across all feature selection methods, particularly when combined with GWO. It achieves the highest accuracy (88.14%) and the lowest RMSE (0.34), highlighting its robustness in handling categorical data and generalizing to unseen instances. While AdaBoost also performs well when paired with GWO, its accuracy and precision are slightly lower than those of CatBoost.

Impact of Feature Selection: GWO significantly enhances classifier performance by effectively selecting relevant features. While PCA and RFE yield moderate results, GA performs similarly to

RFE. These findings underscore the critical role of feature selection in improving the predictive power of ensemble models.

6. Conclusion

Agricultural price forecasting is a complex, interdisciplinary, and evolving research area with significant implications for farmers, traders, and consumers, particularly for perishable crops like vegetables. This study on agricultural commodity price prediction using ensemble learning identifies Grey Wolf Optimization (GWO) and CatBoost as the most effective combination for accurate forecasting. GWO efficiently selects the most relevant features, enhancing model performance, while CatBoost, known for its ability to handle categorical data and strong generalization capabilities, achieves the highest accuracy (88.14%) and lowest RMSE (0.34). Although other classifiers, such as AdaBoost and Gradient Boosting, perform well with GWO, XGBoost consistently underperforms across all feature selection methods. The findings underscore that the GWO-CatBoost combination provides the most reliable and accurate approach for agricultural price forecasting, making it a preferred choice for this predictive task.

References

[1] Weis, A.J., 2007. *The global food economy: The battle for the future of farming*. Zed Books.

[2] Clay, J., 2013. *World agriculture and the environment: a commodity-by-commodity guide to impacts and practices*. Island Press.

[3] Yaşar Dinçer, F.C., Yirmibeşoğlu, G., Narin, M. and Saraç, F.E., 2024. Evaluating the Impact of the COVID-19 Pandemic on the Sustainability of International Trade in Agricultural Products in the Context of Crisis Management: An Assessment of the Agricultural Product Exporting Sectors in Antalya, Türkiye. *Sustainability*, *16*(13), p.5684.

[4] Tripathi, P.K., Singh, C.K., Singh, R. and Deshmukh, A.K., 2023. A farmer-centric agricultural decision support system for market dynamics in a volatile agricultural supply chain. *Benchmarking: An International Journal*, *30*(10), pp.3925-3952.

[5] Sharma, P., Paul, R.K., Meena, D.C. and Anwer, E., 2023. Understanding price volatility and seasonality in agricultural commodities in India. *Agricultural Economics Research Review*, *36*(2), pp.177-188.

[6] Leon, J., 2024. *Forecasting Imported Fruit Prices in the United States Using Neural Networks* (Doctoral dissertation, National University).

[7] Ben Ameur, H., Boubaker, S., Ftiti, Z., Louhichi, W. and Tissaoui, K., 2024. Forecasting commodity prices: empirical evidence using deep learning tools. *Annals of Operations Research*, *339*(1), pp.349-367.

[8] Banerjee, S. and Mondal, A.C., 2023. An ingenious method for estimating future crop prices that emphasises machine learning and deep learning models. *International journal of information technology*, *15*(8), pp.4291-4313.

[9] Effrosynidis, D., Spiliotis, E., Sylaios, G. and Arampatzis, A., 2023. Time series and regression methods for univariate environmental forecasting: An empirical evaluation. *Science of The Total Environment*, *875*, p.162580.

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

[11] Zhang, Y., Liu, J. and Shen, W., 2022. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, *12*(17), p.8654.

[12] Zhang, D., Chen, S., Liwen, L. and Xia, Q., 2020. Forecasting agricultural commodity prices using model selection framework with time series features and forecast horizons. *IEEE access*, *8*, pp.28197-28209.

[13] Bhavani, M. and Mounika, P., 2022. A Novel Model Selection Framework for Forecasting Agricultural Commodity Prices using Time Series Features and Forecast Horizons. *International Journal of Scientific Research in Science and Technology*, pp.134-144.

[14] Gárate-Escamila, A.K., El Hassani, A.H. and Andrès, E., 2020. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, p.100330.

[15] Sya'idah, I.B., Surono, S. and Wen, G.K., 2024. DynamicWeighted Particle Swarm Optimization-Support Vector Machine Optimization in Recursive Feature Elimination Feature Selection. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *23*(3), pp.627-640.

[16] Fang, Y., Yao, Y., Lin, X., Wang, J. and Zhai, H., 2024. A feature selection based on genetic algorithm for intrusion detection of industrial control systems. *Computers & Security*, *139*, p.103675.

[17] Wang, Y., Ran, S. and Wang, G.G., 2024. Role-oriented binary grey wolf optimizer using foraging-following and Lévy flight for feature selection. *Applied Mathematical Modelling*, *126*, pp.310-326.

[18] La Fuente, S., Jennings, E., Lenters, J.D., Verburg, P., Tan, Z., Perroud, M., Janssen, A.B. and Woolway, R.I., 2024. Ensemble modeling of global lake evaporation under climate change. *Journal of Hydrology*, *631*, p.130647.

[19] Bâra, A. and Oprea, S.V., 2024. An ensemble learning method for Bitcoin price prediction based on volatility indicators and trend. *Engineering Applications of Artificial Intelligence*, *133*, p.107991.

[20] Zhang, J. and Chen, X., 2024. A two-stage model for stock price prediction based on variational mode decomposition and ensemble machine learning method. *Soft Computing*, *28*(3), pp.2385-2408.

[21] Abdullah et al., "Intelligent Hybrid ARIMA-NARNET Time Series Model to Forecast Coconut Price," in IEEE Access, vol. 11, pp. 48568-48577, 2023.

[22] Mohanty, M.K., Thakurta, P.K.G. & Kar, S. Agricultural commodity price prediction model: a machine learning framework. *Neural Comput & Applic* **35**, 15109–15128 (2023).

[23] Avinash, G., Ramasubramanian, V., Ray, M., Paul, R.K., Godara, S., Nayak, G.H., Kumar, R.R., Manjunatha, B., Dahiya, S. and Iquebal, M.A., 2024. Hidden Markov guided Deep Learning models for forecasting highly volatile agricultural commodity prices. *Applied Soft Computing*, p.111557.

[24] Zhang, T. and Tang, Z., 2024. Agricultural commodity futures prices prediction based on a new hybrid forecasting model combining quadratic decomposition technology and LSTM model. *Frontiers in Sustainable Food Systems*, *8*, p.1334098.

[25] Rana, H., Farooq, M.U., Kazi, A.K., Baig, M.A. and Akhtar, M.A., 2024. Prediction of Agricultural Commodity Prices using Big Data Framework. *Engineering, Technology & Applied Science Research*, *14*(1), pp.12652-12658.

[26] Sun, F., Meng, X., Zhang, H., Wang, Y. and Liu, P., 2024. Prediction of Weekly Price Trend of Garlic Based on Classification Algorithm and Combined Features. *Horticulturae*, *10*(4), p.347.

[27] Saini, P. and Nagpal, B., 2024. Analysis of missing data and comparing the accuracy of imputation methods using wheat crop data. *Multimedia Tools and Applications*, *83*(14), pp.40393-40414.

[28] Bukhari, S.B., 2024. Crop recommendation system using machine learning. a data-driven approach to sustainable agriculture. *Journal of Techno Trainers*, *1*(2), pp.26-35.

[29] Sachdeva, R.K., Bathla, P., Rani, P., Kukreja, V. and Ahuja, R., 2022, April. A systematic method for breast cancer classification using RFE feature selection. In *2022 2nd international conference on advance computing and innovative technologies in engineering (ICACITE)* (pp. 1673-1676). IEEE.

[30] Altarabichi, M.G., Nowaczyk, S., Pashami, S. and Sheikholharam Mashhadi, P., 2023, July. Fast Genetic Algorithm for feature selection—A qualitative approximation approach. In *Proceedings of the companion conference on genetic and evolutionary computation* (pp. 11-12).

[31] Mirjalili, S., Mirjalili, S.M. and Lewis, A., 2014. Grey wolf optimizer. *Advances in engineering software*, *69*, pp.46-61.

[32] V. Bobkov, A. Bobkova, S. Porshnev and V. Zuzin, "The application of ensemble learning for delineation of the left ventricle on echocardiographic records," 2016 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, 2016, pp. 1-5.

[33] X. Shu and P. Wang, "An Improved Adaboost Algorithm Based on Uncertain Functions," 2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, Wuhan, 2015, pp. 136-139

[34] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 2018, 31, 6638–6648.

[35] Liu, H.; Guo, L.; Li, H.; Zhang, W.; Bai, X. Matching areal entities with CatBoost ensemble method. Geogr. Inf. Sci. 2022, 24, 2198–2211.