Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Hybrid Grey Butterfly Optimizer-Based Feature Selection for Enhancing Agricultural Commodity Price Prediction Using Machine Learning Classifiers

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

[1]Research scholar, Department of computer science Ppg arts and science,

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

Abstract

One of the most important tasks for economic planning and decision-making in the agriculture industry is the prediction of agricultural commodity prices. However, high dimensionality and redundancy are common in this field's datasets, increasing computing cost and decreasing prediction accuracy. To address these problems, this study proposes a novel Hybrid Grey Butterfly Optimizer (HGBO) for feature selection that combines the exploratory nature of the Butterfly Optimization Algorithm (BOA) with the leadership-driven search capabilities of Grey Wolf Optimization (GWO). The proposed hybrid approach enhances the feature selection process by avoiding local optima and preventing premature convergence. The selected feature subsets were assessed using machine learning classifiers including XGBoost, Random Forest, Gradient Boosting, and Support Vector Machines (SVM) to know the efficacy of the HGBO strategy. The performance of the proposed method was assessed in comparison to alternative optimization techniques, utilizing a range of evaluation metrics such as accuracy, precision, recall, and the F1-score. The findings indicate that HGBO demonstrates superior performance by identifying the most pertinent feature subsets, thereby enhancing predictive accuracy. This research underscores the HGBO algorithm's capacity for forecasting agricultural prices, presenting a valuable approach for addressing high-dimensional data challenges within agricultural analytics. .

Keywords: Price prediction, Grey Wolf Optimization, machine learning, Butterfly Optimization Algorithm

Introduction

The prediction of agricultural commodity prices (ACPP) is of paramount importance for economic planning, the formulation of effective policies, and the maintenance of market stability [1-3]. Precise price forecasts can empower farmers, retailers, and regulatory bodies to make informed decisions, thereby contributing to food security and improving supply chain management [4]. The inherent complexity of agricultural data, which is characterized by high dimensionality, redundancy, and noise, significantly hinders predictive modeling [5-7]. Traditional forecasting techniques sometimes encounter inefficiencies while handling such data, which lead to increased processing load and decreased model accuracy [8–10].

Feature selection is a critical step in addressing these problems by identifying the most beneficial characteristics and eliminating redundant or superfluous features [11]. Conventional feature selection techniques, such statistical and filter-based methods, may not adequately capture complex relationships between variables. Consequently, metaheuristic optimization techniques have emerged as useful tools for enhancing feature selection in machine learning-based agricultural price prediction [12].

The Hybrid Grey Butterfly Optimizer (HGBO), which combines the Butterfly Optimization Algorithm (BOA) and Grey Wolf Optimization (GWO), is a novel feature selection technique presented in this research [13–15]. This hybridization is motivated by the complementary benefits of these two methods. While GWO is well known for its leadership-driven hierarchical search mechanism that successfully strikes a balance between exploration and exploitation, BOA enhances exploratory skills by mimicking butterfly foraging behavior using fragrance-based movement strategies. By combining these two approaches, HGBO aims to enhance feature selection performance, boost search efficiency, and get beyond local optima trapping.

To evaluate its effectiveness, we used the HGBO-based feature selection method with machine learning classifiers including XGBoost, Random Forest (RF), Gradient Boosting (GB), and Support Vector Machines (SVM). These classifiers were chosen because of their adaptability and resilience to complex agricultural datasets. Common classification metrics, including accuracy, precision, recall, and F1-score, were used to assess HGBO's performance and compare it to other optimization techniques now in use.

Although metaheuristic algorithms are widely used in predictive modeling, their use in high-dimensional agricultural datasets is still missing. Most recent studies employ single-metaheuristic methods, such as the traditional Grey Wolf Optimizer (GWO) or the Butterfly Optimization Algorithm (BOA). However, when applied to volatile agricultural price data, which are marked by high multi-collinearity and seasonal noise, these single algorithms frequently suffer from a lack of variation in their search area. Specifically, when engaging with the huge feature sets that define modern rural economy, the regular BOA may exhibit slow exploitation phases, whereas GWO often exhibits early convergence into local optima due to its rigid leadership structure.

There isn't a complete model that balances the strong, leadership-driven exploitation of GWO with the stochastic, fragrance-based exploration of BOA. This study closes this gap and ensures a more relevant and condensed feature subset than is currently achievable with independent optimization algorithms by developing hybrid architecture (HGBO) to manage the non-linear complexity of agricultural commodity movements. The primary contributions of this research are as follows:

Development of a Novel Hybrid Architecture: By successfully fusing the hierarchical exploitation of Grey Wolf Optimization (GWO) with the fragrance-based global exploration of the Butterfly Optimization Algorithm (BOA), the Hybrid Grey Butterfly Optimizer (HGBO) removes local optima trapping in non-linear datasets.

Enhanced Feature Selection Logic: introducing a multi-strategy search mechanism to efficiently finds a compact, high-relevance feature subset from high-dimensional agricultural data.

Empirical Validation on Volatile Data: Support Vector Machines (SVM) were determined to be the most successful predictive basis for agricultural commodity price forecasting after a detailed comparison of the HGBO framework with various state-of-the-art machine learning classifiers (XGBoost, RF, GB, and SVM).

Computational Efficiency: demonstrating that the hybridized method reduces processing costs by eliminating redundant data features, outperforming traditional single-metaheuristic techniques.

The experimental results demonstrate that the proposed HGBO framework significantly improves prediction accuracy by selecting a limited and highly relevant subset of attributes. The hybrid approach effectively reduces computational complexity while enhancing model interpretability, making it a feasible choice for high-dimensional agricultural datasets. The findings show that HGBO has the potential to be a useful tool for agricultural analytics, facilitating better decision-making and more precise price forecasts.

The following sections of this article are arranged as follows: Section 2 reviews related studies on agricultural price prediction and feature selection. The methodology, the HGBO algorithm, and the machine learning classifiers used for evaluation are all fully explained in Section 3. Section 4 presents the experimental setup and results, while

Section 5 provides a commentary. In Section 6, the study is eventually ended with suggestions for more investigation.

Literature Review

In the realm of agricultural commodity price prediction, the combination of machine learning techniques with metaheuristic optimization algorithms has attracted a lot of attention in recent years. By efficiently choosing pertinent characteristics and fine-tuning model parameters, these methods seek to improve prediction accuracy.

Weng et al. [16] used the AutoRegressive Integrated Moving Average (ARIMA) model, the back propagation (BP) network technique, and the recurrent neural network (RNN) method to predict the short- and long-term prices of horticulture goods such as eggplant, tomato, and cucumber. Small-scale periodic data is a good fit for ARIMA, while daily data is not.

Neural network methods, including BP network and RNN, can predict daily, weekly, and monthly price trends, making them more suitable for large-scale data. Deep learning methods are expected to become the mainstream method for agricultural product price forecasting.

Li et al [17] suggests a multivariate approach to financial time-series forecasting. Our approach incorporates validated noise injection and upgraded deep neural networks to solve the multimodal, non-stationarity, and long- and short-term aspects of multivariate time-series. Long- and short-term recurrent neural networks are used for multivariate time-series forecasting, multimodal variational autoencoders are used to extract deep high-level features, and a certified noise injection mechanism inspired by differential privacy is proposed to increase prediction accuracy and robustness.

Anurag Tiwari [18] enhances the Butterfly Optimization Algorithm's local search capability to present a better and more computationally efficient version of the traditional BBOA. Using three distinct transfer functions (S, U, and V-shaped), twelve binary variations were first created. The quality of each solution is assessed based on its corresponding fitness function scores. Next, we investigated BOA's local search capability using Adaptive β−Hill Climbing, another recently created optimization approach, to calculate high-quality solutions. This optimization method uses two stochastic operators, the β-operator (mutation operator) and the N-operator (neighborhood operator), to generate better offspring than parent solutions.

Binary variants of the Butterfly Optimization Algorithm (BOA) are presented by Sankalap Arora and Priyanka Anand [19] to choose the optimal feature subset for classification in a wrapper mode. The two proposed binary versions of BOA are used to select the optimal feature combination that reduces the number of characteristics needed while maximizing classification accuracy. The study shown that this hybrid strategy increases the accuracy of agricultural commodity price predictions by effectively balancing exploration and exploitation throughout the search phase.

The development of a system for forecasting volatility in the agricultural commodities trading arena is described by Le et al. [20]. The system accepts raw financial data as input and produces trading decision-support using a volatility prediction based on the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and Long Short-term Memory (LSTM) models. Increasing total profitability through efficient trading loss management is the system's main goal. Additionally, a denoising technique is used to improve overall performance and reduce the effect of market noise.

Yang [21] introduces a novel machine learning ensemble strategy that combines deconstruction algorithms with physical optimization techniques for commodities futures price prediction. The best modal decomposition results are first obtained by optimizing the VMD (variational mode decomposition) using the RIME algorithm. Then, using the

ELM (Extreme Learning Machines) and FA (Fourier Attention) models, respectively, the trend and seasonal terms are predicted, and the results are finally synthesized.

Fard [22] introduces a multi-objective optimization approach for sustainable harvest planning, utilizing fuzzy logic and multiple objectives to minimize greenhouse gas emissions and waste. It also proposes a revised version of the non-dominated sorting genetic algorithm (NSGEA) to address the complexity of large-scale networks, focusing on the blueberry industry in Canada.

Liu et al [23] presents an EEMD-NAGU hybrid prediction model for soybean futures price, utilizing Ensemble Empirical Mode Decomposition (EEMD) and New Attention Gate Unit (NAGU). The model processes soybean futures price data into multiple IMFs and residual sequences, calculates sample entropy, and reconstructs the IMF into low-frequency, medium-frequency, and high-frequency components. NAGU improves learning capability by capturing historical data and reducing noise.

Proposed Model

Figure 1 shows the various steps of the Hybrid Grey Butterfly Optimizer (HGBO)-Based Feature Selection for Agricultural Commodity Price Prediction design. Data preparation, which includes managing missing values, normalization, and feature scaling, is first applied to time-series agricultural commodity price data. The most pertinent characteristics are then extracted using the HGBO-based feature selection approach, which combines the advantages of Butterfly Optimization Algorithm (BOA) and Grey Wolf Optimization (GWO) to increase search efficiency and prevent local optima.

For price prediction, machine learning classifiers such as XGBoost, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) are fed the optimum feature subset. To identify the top-performing model, the trained models are assessed using common performance metrics including accuracy, precision, recall, and F1-score. In order to ensure that the chosen model offers reliable and accurate agricultural price forecasts and supports well-informed decision-making for economic planning and market stability, an analysis phase is finally carried out to interpret the results.

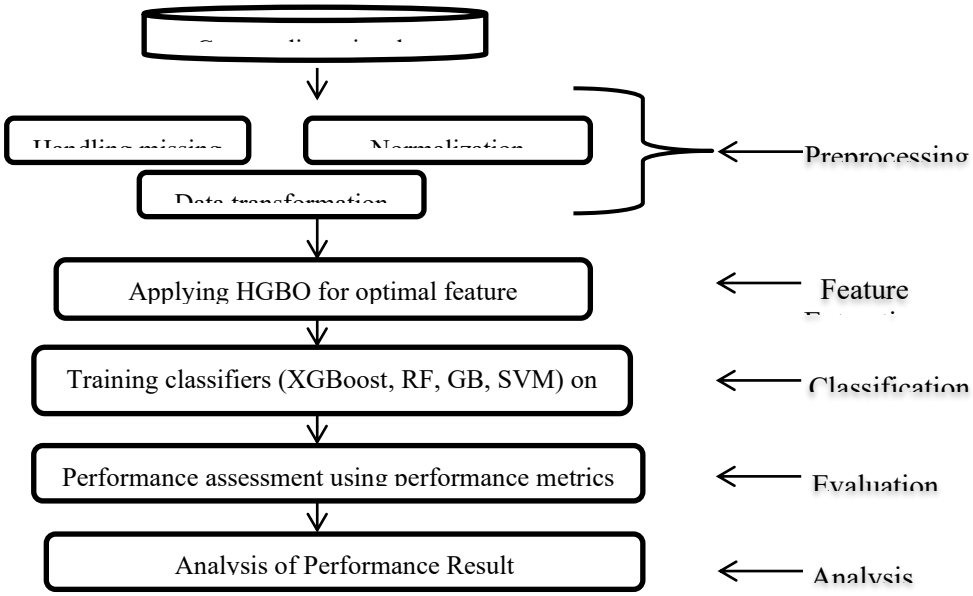

Figure 1. Proposed HGBO based ACPP architecture

Dataset Description

The dataset employed in this study was sourced from a publicly accessible agricultural market repository hosted by the India Data Portal, which provides daily commodity-wise price and arrival statistics collected from Agricultural Produce Market Committees (APMCs) across various regions of India. To capture diverse market behaviors and pricing patterns, five commonly traded vegetables—onion, carrot, brinjal, potato, and tomato—were selected for analysis. Each commodity dataset was initially obtained as an independent comma-separated values (CSV) file, comprising attributes such as district, commodity category, commodity name, variety, modal price, price unit, transaction date, etc. In order to facilitate unified multi-commodity modeling, all separate datasets were combined into a single integrated dataset after data collecting using schema harmonization and data cleaning techniques.

For price prediction in classification tasks, a continuous modal price was converted into a categorical target variable through a percentile-based discretization method. The 33rd and 66th percentiles set thresholds that divide the data into three classes: Low (below the 33rd percentile), Medium (between the 33rd and 66th percentiles), and High (above the 66th percentile). This approach ensures clear class differentiation and balanced distributions, which is advantageous for supervised learning in predicting price ranges.

Preprocessing

In order to ensure that the dataset is clean, organized, and appropriate for machine learning models, data preparation is an essential stage in the prediction of agricultural commodity prices. Three essential elements are involved in this phase: data transformation, normalization, and addressing missing values.

Handling Missing Values

Sensor malfunctions, human mistake, or inconsistent data recording are the causes of missing values in agricultural datasets. There are many methods for dealing with missing data:
In agricultural commodity price prediction, where data gaps might arise from sensor failures, missing records, or inconsistent data collecting, mean imputation is a frequently used approach to address missing values in datasets. The mean (average) of the observed values for that characteristic is used to replace missing values.

Let $Q = q_1, q_2, q_3, \cdots q_n$ the mean imputation is calculates by following equation.

$$\hat{Q}_x = \frac{1}{N}\sum_{y=1}^{N} q_y \tag{1}$$

where $\hat{Q}_x$ is the imputed value, q_y represents non-missing values, N is the total count of available (non-missing) values.

Normalization

By putting numerical values within a common range, normalization is crucial for guaranteeing that each feature contributes equally to the model. Typical methods for normalizing consist of:
A feature scaling approach called Min-Max Scaling [24] is used to convert numerical data into a defined range, usually [0, 1] or [-1, 1]. In order to normalize factors like soil qualities, weather, and market prices and prevent features with varying magnitudes from dominating machine learning models, this approach is frequently employed in agricultural commodity price prediction and yield estimate. The following is the computation formula.

$$Q' = \frac{Q - Q_{min}}{Q_{max} - Q_{min}} \tag{2}$$

Q stands for the initial value, Q_{max} and Q_{min} for the feature's minimum and maximum values, and Q'scaled value in the interval [0, 1].

If the dataset has been appropriately preprocessed, the Hybrid Grey Butterfly Optimizer (HGBO) may effectively choose relevant features while minimizing redundancy. Proper treatment of missing data, normalization, and transformation not only increases model accuracy but also reduces computer complexity, leading to more accurate and dependable agricultural commodity price projections.

Data Transformation Using Principal Component Analysis (PCA)

In machine learning and predictive modeling, data transformation is an essential preprocessing step, especially for high-dimensional datasets. One of the most popular methods for reducing dimensionality is Principal Component Analysis (PCA) [25]. It preserves as much variation as feasible while reducing a high-dimensional dataset to a smaller collection of uncorrelated variables known as principle components (PCs).

Datasets for predicting agricultural commodity prices frequently include a variety of characteristics, including market movements, soil characteristics, and climatic variables, all of which may be strongly connected.

PCA helps by reducing redundancy, improving computational efficiency, and enhancing model performance.

Table 1: PCA data with two principal components (PC1 and PC2)

	PC1	**PC2**
0	4.8916	0.003
1	-0.5644	-0.0732
2	-0.4941	-0.0318
3	-0.6697	-0.0357
4	-0.586	-0.0222
5	-0.6689	0.0794
6	-0.6463	0.1563
7	-0.7115	-0.0232

This pseudocode outlines the key steps of PCA, from preprocessing to transformation, ensuring dimensionality reduction while retaining important information.

Table 2: Pseudocode for Data Transformation Using Principal Component Analysis (PCA)

Algorithm: Principal Component Analysis (PCA)
Input: Dataset (X) with n samples and m features **Output:** Transformed dataset (X_PCA) with reduced dimensions Step 1: Standardize the dataset: FOR each feature in X: Compute mean (μ) and standard deviation (σ)

```
        Standardized_feature = (Feature - μ) / σ
Step 2: Compute Covariance Matrix
      Compute the covariance matrix C:
      C = (1 / (n-1)) * (X_standardized^T * X_standardized)
Step 3: Compute Eigenvalues and Eigenvectors
     Solve for eigenvalues (λ) and eigenvectors (V) of C:
        [V, λ] = Eigen_Decomposition(C)
     Sort eigenvalues in descending order
     Select top k eigenvectors corresponding to the largest k eigenvalues
Step 4: Transform Data
     Compute the transformed dataset:
       X_PCA = X_standardized * V_k   // Project data onto k principal components
Step 5: Evaluate and Interpret Results
    Compute explained variance ratio:
       Explained_Variance = λ_k / sum(λ)
    Visualize explained variance to determine optimal k
    Return X_PCA (transformed data)
End Algorithm
```

Hybrid Grey Butterfly Optimizer (HGBO) - Feature Selection

The Hybrid Grey Butterfly Optimizer (HGBO) is a novel feature selection algorithm that combines the Grey Wolf Optimization (GWO) and Butterfly Optimization Algorithm (BOA) to enhance the selection of optimal features in high-dimensional datasets. While BOA improves exploration through fragrance-based adaptive movements, GWO uses a leadership-driven search mechanism that balances exploration and exploitation by mimicking grey wolf hunting behavior. HGBO successfully avoids local optima, increases search efficiency, and delays premature convergence by combining these complementing techniques. By removing unnecessary characteristics and choosing the most pertinent ones, this hybrid method lowers computational complexity and improves model accuracy. HGBO guarantees that only the most useful aspects are included in forecasting models for agricultural commodity prices, resulting in more accurate and understandable forecasts.

Steps Involved in Hybrid Grey Butterfly Optimizer (HGBO)

Step 1: Initialize Population

In the HGBO initialization step, a variety of candidate feature subsets are created, each of which is represented as a binary or continuous vector that corresponds to certain characteristics. Because the population size is predetermined, exploration and computational efficiency are balanced. Every potential solution is started at random, with a vector value of 0 denoting feature exclusion and 1 denoting feature selection. Furthermore, parameters that influence the Butterfly Optimization Algorithm (BOA) [29–31] and Grey Wolf Optimization (GWO) [26–28], such as the control coefficients for exploration-exploitation balance, are initialized. In order to avoid premature convergence and

guarantee thorough search coverage throughout the feature space, the starting population's diversity is essential. Machine learning classifiers are then used to assess each candidate subset's fitness, laying the groundwork for iterative optimization utilizing the hybridized method of GWO and BOA.

Step 2: Evaluate Fitness

The fitness of each candidate subset is then evaluated using machine learning classifiers, setting the stage for iterative optimization using the hybridized approach of GWO and BOA.

Mathematically, the fitness function can be formulated as:

$$F = \alpha \times Accuracy - \beta \times \frac{|S|}{|T|}$$

F is the fitness value; S is the number of selected features; T is the total number of accessible features; α and β are weight factors that balance accuracy and feature subset size; and Accuracy is the classification accuracy achieved using the chosen feature subset.

Step 3: Apply Grey Wolf Optimization (GWO) Search Strategy

The Grey Wolf Optimization (GWO) algorithm is inspired by the hierarchical leadership and hunting strategies of grey wolves in nature. In Hybrid Grey Butterfly Optimizer (HGBO), GWO is incorporated to enhance exploitation by guiding the feature selection process efficiently. The GWO search strategy follows four main phases: encircling prey, hunting, attacking, and convergence.

Encircling Prey

Grey wolves encircle their prey during hunting. In GWO, this behavior is mathematically modeled to update the positions of feature subsets. The equations governing this step are:

$$R = |F.Q_b - Q|$$

$$Q_{new} = Q_b - E.F$$

Where Q_b refers the position of best solution, Q denotes the position of current search agent, R represents distance between the search agent and the best solution, $E\ and\ F$ are coefficient vectors, controlling exploration and exploitation, defined as

$$E = 2e.x_1 - e\ , F = 2.x_2$$

where e decreases linearly from 2 to 0 over iterations, and x_1, x_2 are random numbers in [0,1] to maintain randomness.

Hunting

The GWO hunting process is driven by a hierarchical leadership structure consisting of: $\alpha, \beta\ and\ \gamma$ represent best feature subset, the second-best subset, assisting in the search and the third-best subset, providing guidance respectively.

The remaining search agents (Omega wolves) update their positions relative to the top three leaders using:

$$R_\alpha = |F_1.Q_\alpha - Q|$$

$$R_\beta = |F_2.Q_\beta - Q|$$

$$R_\gamma = |F_3.Q_\gamma - Q|$$

$$Q_{new} = \frac{Q_1 + Q_2 + Q_3}{3}$$

where Q_1, Q_2, Q_3 are updated positions based on α, β, and δ wolves. This mechanism ensures the search agents move toward promising feature subsets.

Attacking Prey

The shift from exploration (global search) to exploitation (local refinement) is controlled by

The transition from exploration (global search) to exploitation (local refinement) is controlled by $\boldsymbol{E}$, which decreases over iterations. When$|E| < 1$, wolves intensify their search by moving closer to the best feature subset, enabling fine-tuned selection.

Convergence and Stopping Criteria

An ideal balance between exploration and exploitation is ensured by the convergence and halting criteria of the Grey Wolf Optimization (GWO) Search Strategy in the Hybrid Grey Butterfly Optimizer (HGBO). The hierarchical leadership system, in which the alpha, beta, and delta wolves constantly adjust their locations in response to the movement of the prey (optimal solution), directs the convergence of GWO. Until the fitness value stabilizes within a predetermined threshold or the maximum number of iterations is achieved, the algorithm iteratively improves the solutions. Conditions like reaching an acceptable error margin, surpassing a predetermined computational budget, or showing little gain in fitness over subsequent rounds are examples of halting criteria.

Step 4: Apply Butterfly Optimization Algorithm (BOA) for Exploration

The BOA step ensures that unexplored regions of the search space are adequately examined. It helps in finding new potential feature subsets by continuously updating the butterfly positions based on fragrance levels. BOA's global search strategy introduces randomness, ensuring diverse feature combinations are considered. This step complements the Grey Wolf Optimization (GWO) phase, as it prevents premature convergence and maintains a diverse set of feature subsets for selection.

Butterflies follow the Fragrance-Based Search where each butterfly emits a fragrance that influences its movement. The definition of the fragrance is given in following equation

$$F_x = sI_x^c$$

Where F_x is the fragrance of the x^{th} butterfly, s is a sensory modality constant, I_x is the stimulus intensity (fitness value), c is the power exponent controlling fragrance perception.

Global search (Exploration)

Butterflies move towards the best-known solution, ensuring that the algorithm explores promising areas efficiently defined as global search (Exploration):

$$Q_x^{t+1} = Q_x^t + r \times F_x \times (Q^* - Q_x^t)$$

Local Search (Exploitation):

If the probability condition is met, a butterfly updates its position relative to another randomly chosen butterfly:

$$Q_x^{t+1} = Q_x^t + r \times F_x \times \left(Q_y^t - Q_k^t\right)$$

Where $Q_y^t - Q_k^t$ are the positions of two randomly selected butterflies.

Step 5: Hybridization and Adaptive Mechanism

The Hybridization and Adaptive Mechanism in HGBO combines the strengths of both optimization approaches to enhance feature selection efficiency. Hybridization integrates the structured leadership-based exploration of GWO with the adaptive and fragrance-driven movement of BOA, ensuring a balance between global exploration and local exploitation. The adaptive mechanism dynamically adjusts key parameters, such as the influence of leader wolves in GWO and the fragrance factor in BOA, based on iteration progress. This adaptation prevents premature convergence and enhances solution diversity. Additionally, mutation is introduced to refine the feature subset by flipping bits probabilistically, ensuring the optimizer explores new potential solutions while maintaining strong convergence toward an optimal feature selection set.

Step 6: Selection of Optimal Feature Subset

- Rank solutions based on fitness and retain the best-performing feature subset. The selection of the optimal feature subset in HGBO is driven by a balance of exploration and exploitation, ensuring the best features are chosen for model performance enhancement.

Step 7: Termination Criteria

- Repeat steps 3 to 6 until the stopping condition is met (e.g., maximum iterations or no significant improvement).
- Output the optimal feature subset for further analysis or model training.

This hybrid approach enhances feature selection efficiency, improves model interpretability, and ensures better predictive performance in agricultural commodity price forecasting.

Pseudocode for Hybrid Grey Butterfly Optimizer (HGBO)

```
df = pd.read_csv("agriculture_dataset.csv")
df.fillna(df.median(numeric_only=True), inplace=True)
label_encoders = {}
for col in df.select_dtypes(include=['object']).columns:
    label_encoders[col] = LabelEncoder()
    df[col] = label_encoders[col].fit_transform(df[col])
scaler = MinMaxScaler()
df[df.columns] = scaler.fit_transform(df[df.columns])
X = df.drop(columns=['Modal Price'])  # Features
y = df['Modal Price']  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# HGBO Feature Selection
def initialize_population(size, dim):
```

```
    return np.random.rand(size, dim)
def fitness_function(solution, X_train, X_test, y_train, y_test):
    selected_features = [i for i in range(len(solution)) if solution[i] == 1]
    if len(selected_features) == 0:
        return 0  # Avoid empty feature sets
        X_train_fs = X_train[:, selected_features]
    X_test_fs = X_test[:, selected_features]
    model = RandomForestClassifier()
    model.fit(X_train_fs, y_train)
    y_pred = model.predict(X_test_fs)
    return accuracy_score(y_test, y_pred)
initialize_population(pop_size, feature_dim):
    return np.random.randint(2, size=(pop_size, feature_dim))
gwo_update(wolves, alpha, beta, delta, a):
    new_positions = np.zeros_like(wolves)
    for i in range(wolves.shape[0]):
        A1, A2, A3 = a * (2 * np.random.rand() - 1), a * (2 * np.random.rand() - 1), a * (2 * np.random.rand() - 1)
        C1, C2, C3 = 2 * np.random.rand(), 2 * np.random.rand(), 2 * np.random.rand()
        D_alpha = abs(C1 * alpha - wolves[i])
        D_beta = abs(C2 * beta - wolves[i])
        D_delta = abs(C3 * delta - wolves[i])
        X1 = alpha - A1 * D_alpha
        X2 = beta - A2 * D_beta
        X3 = delta - A3 * D_delta
        new_positions[i] = (X1 + X2 + X3) / 3
    return (new_positions > 0.5).astype(int)
boa_update(butterflies, best_solution, fragrance_factor=0.1):
    new_positions = np.copy(butterflies)
    for i in range(butterflies.shape[0]):
        if random.random() < 0.5:
            step = fragrance_factor * (best_solution - butterflies[i])
        else:
```

```
            step = fragrance_factor * (butterflies[random.randint(0, len(butterflies) - 1)] - butterflies[i])
        new_positions[i] = butterflies[i] + step
    return (new_positions > 0.5).astype(int)
hybrid_mutation(population, mutation_rate=0.1):
    new_population = np.copy(population)
    for i in range(len(population)):
        if random.random() < mutation_rate:
            mutate_index = random.randint(0, population.shape[1] - 1)
            new_population[i][mutate_index] = 1 - new_population[i][mutate_index]
    return new_population
HGBO(X, y, pop_size=50, max_iter=100):
    feature_dim = X.shape[1]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
    population = initialize_population(pop_size, feature_dim)
    fitness_scores = np.array([fitness_function(ind, X_train, X_test, y_train, y_test) for ind in population])
    alpha, beta, delta = population[np.argsort(fitness_scores)[-3:]]  # Top 3 solutions
    for iter in range(max_iter):
        a = 2 - iter * (2 / max_iter)  # Linearly decreasing factor for GWO
        gwo_population = gwo_update(population, alpha, beta, delta, a)
        boa_population = boa_update(population, alpha)
        hybrid_population = hybrid_mutation((gwo_population + boa_population) // 2)
        fitness_scores = np.array([fitness_function(ind, X_train, X_test, y_train, y_test) for ind in hybrid_population])
        best_idx = np.argmax(fitness_scores)
        if fitness_scores[best_idx] > fitness_function(alpha, X_train, X_test, y_train, y_test):
            alpha = hybrid_population[best_idx]
        print(f'Iteration {iter+1}: Best Accuracy = {fitness_scores[best_idx]:.4f}')
    return alpha, fitness_scores[best_idx]
```

The Hybrid Grey Butterfly Optimizer (HGBO) is designed to enhance feature selection by integrating GWO and BOA and their procedure is given in table 2. This hybrid approach aims to optimize the selection of features from an agriculture dataset to improve predictive accuracy. The process starts with data preprocessing, including handling missing values using median imputation, encoding categorical variables with Label Encoding, and applying Min-Max Scaling to normalize numerical features. The dataset is then split into training and testing sets for model evaluation.

The optimization begins with the initialization of a binary population, where each individual represents a feature selection subset. The fitness function is defined using a Random Forest classifier, evaluating the accuracy of selected features. The GWO component updates feature selection using the leadership hierarchy of alpha, beta, and delta wolves, dynamically adjusting positions based on adaptive coefficients. Simultaneously, BOA introduces diversity, guiding feature movement based on fragrance factors and adaptive position updates. A hybridization step combines the best features of GWO and BOA, ensuring optimal feature selection. Additionally, mutation is applied to prevent premature convergence by randomly flipping feature selection bits.

Throughout iterations, the best solutions are retained, and convergence is monitored based on fitness improvement. The algorithm iterates until a stopping criterion is met, such as reaching the maximum number of iterations or stabilizing the best accuracy. The final selected feature subset represents the most informative attributes for the prediction model, maximizing accuracy while reducing dimensionality. This HGBO-based feature selection improves model efficiency, ensuring better generalization on real-world agricultural datasets.

Result and Discussion

This study implemented Hybrid Grey Butterfly Optimizer (HGBO)- based Feature Selection for Agricultural Commodity Price Prediction using machine learning classifiers. The entire implementation was conducted in Python, utilizing essential libraries such as NumPy, Pandas, Scikit-learn, and Matplotlib for data processing, model training, and visualization. The dataset used for agricultural commodity price prediction underwent multiple preprocessing steps to enhance model performance. The dataset was collected from kaggle data repository and divided into 80% training data and 20% testing data to evaluate model performance effectively.

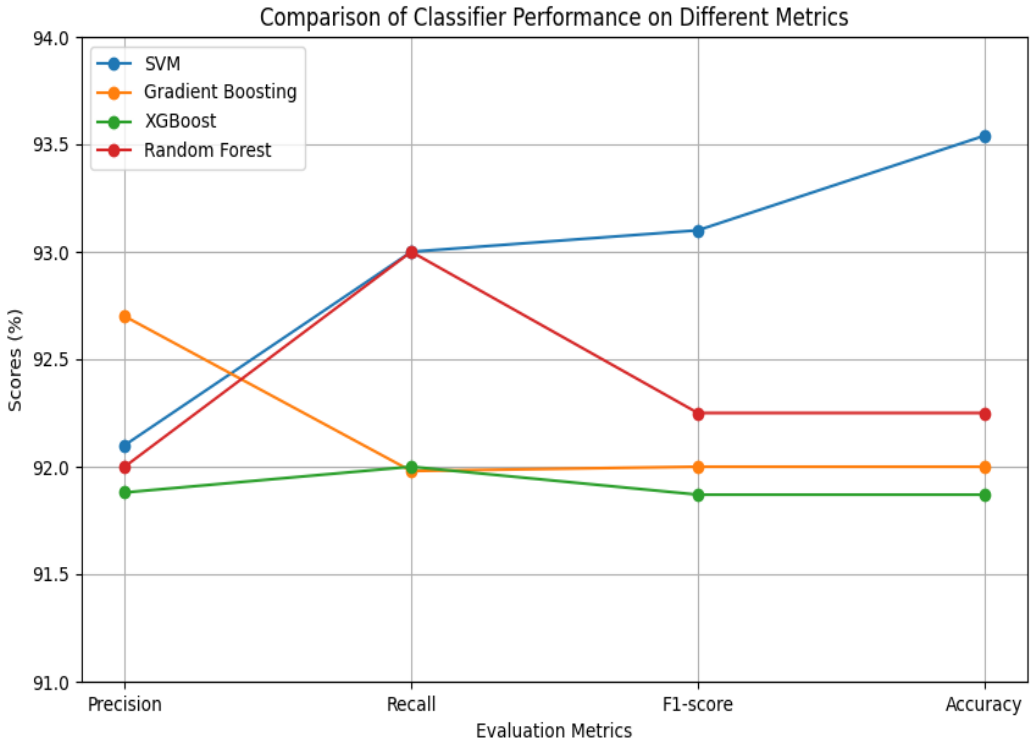

Figure 2. Performance comparison of different metrics

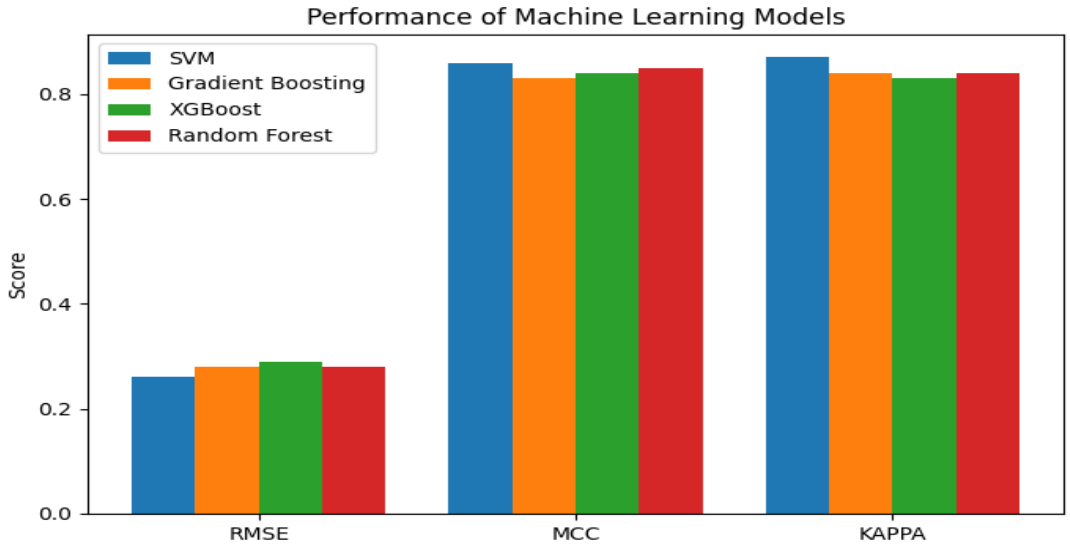

Figure 3. RMSE, MCC and KAPPA value comparison with different approach

The performance evaluation of agricultural commodity price prediction using different machine learning classifiers—SVM, Gradient Boosting, XGBoost, and Random Forest—was conducted based on four key metrics: Precision, Recall, F1-score, and Accuracy.

Among these classifiers, Support Vector Machine (SVM) achieved the highest accuracy of 93.54%, making it the most effective model for price prediction. It also attained the best F1-score (93.10%), indicating a well-balanced performance between precision and recall. The high Recall (93.00%) suggests that SVM effectively captured most relevant patterns in the dataset, while its Precision (92.10%) signifies that false positives were minimized. Gradient Boosting followed closely with an accuracy of 92.00%, showing strong performance with a Precision of 92.70% but slightly lower Recall (91.98%), which indicates that while its predictions were accurate, it may have missed some actual instances. XGBoost, with an accuracy of 91.87%, demonstrated robust performance but lagged behind due to its slightly lower recall and F1-score. Random Forest performed comparably to Gradient Boosting with an accuracy of 92.25%, but it did not outperform SVM. Overall, these results highlight that SVM is the most suitable model for predicting agricultural commodity prices in this scenario, as it consistently outperformed other models across all metrics, making it the optimal choice for reliable and accurate predictions.

Conclusion

A Hybrid Grey Butterfly Optimizer (HGBO)-based feature selection framework is proposed to enhance agricultural commodity price prediction using machine learning classifiers. The integration of HGBO effectively selected the most relevant features, reducing dimensionality while preserving critical information for accurate predictions. The preprocessing phase involved handling missing values, encoding categorical data, and applying Min-Max Scaling to standardize the dataset, ensuring optimal input for the classification models. The selected feature subset was evaluated using Support Vector Machine (SVM), Gradient Boosting, XGBoost, and Random Forest classifiers. Among these, SVM achieved the highest accuracy of 93.54%, demonstrating its superior capability in capturing complex patterns within the agricultural price data. The results highlight the effectiveness of HGBO-based feature selection in improving classification accuracy and computational efficiency in agricultural commodity price prediction. This approach can serve as a valuable tool for policymakers, farmers, and market analysts to make informed decisions regarding commodity pricing. Future work can focus on extending this framework to real-time price forecasting and integrating deep learning models to further enhance prediction accuracy.

References

[1] N. -B. -v. Le, Y. -S. Seo and J. -H. Huh, "AgTech: Volatility Prediction for Agricultural Commodity Exchange Trading Applied Deep Learning," in IEEE Access, vol. 12, pp. 153898-153910, 2024.

[2] Silva, A., Caraiani, P., Miranda-Pinto, J. and Olaya-Agudelo, J., 2024. Commodity prices and production networks in small open economies. Journal of Economic Dynamics and Control, Elsevier, 168, p.104968.

[3] H. Li, Y. Cui, S. Wang, J. Liu, J. Qin and Y. Yang, "Multivariate Financial Time-Series Prediction With Certified Robustness," in *IEEE Access*, vol. 8, pp. 109133-109143, 2020

[4] Kalimuthu, T., Kalpana, P., Kuppusamy, S. and Sreedharan, V.R., 2024. Intelligent decision-making framework for agriculture supply chain in emerging economies: Research opportunities and challenges. *Computers and Electronics in Agriculture*, *219*, p.108766.

[5] Kumar, R., Bhanu, M., Mendes-Moreira, J. and Chandra, J., 2024. Spatio-Temporal Predictive Modeling Techniques for Different Domains: a Survey. *ACM Computing Surveys*, *57*(2), pp.1-42.

[6] Hansen, J.W., 2002. Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges. *Agricultural systems*, Elsevier, *74*(3), pp.309-330.

[7] Wu, T., Gao, X., An, F., Sun, X., An, H., Su, Z., Gupta, S., Gao, J. and Kurths, J., 2024. Predicting multiple observations in complex systems through low-dimensional embeddings. *Nature Communications*, Elsevier, *15*(1), p.2242.

[8] Cortez, C.T., Saydam, S., Coulton, J. and Sammut, C., 2018. Alternative techniques for forecasting mineral commodity prices. *International Journal of Mining Science and Technology*, Elsevier, *28*(2), pp.309-322.

[9] Shaikh, T.A., Mir, W.A., Rasool, T. and Sofi, S., 2022. Machine learning for smart agriculture and precision farming: towards making the fields talk. *Archives of Computational Methods in Engineering*, springer, *29*(7), pp.4557-4597.

[10] Tantalaki, N., Souravlas, S. and Roumeliotis, M., 2019. Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of agricultural & food information*, *20*(4), pp.344-380.

[11] Raja, S.P., Sawicka, B., Stamenkovic, Z. and Mariammal, G., 2022. Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. *IEEE Access*, *10*, pp.23625-23641.

[12] Sulaiman, M.H., Mustaffa, Z., Saari, M.M. and Abas, M.F., 2024. Wind power forecasting with metaheuristic-based feature selection and neural networks. *Cleaner Energy Systems*, *9*, p.100149.

[13] Makhadmeh, S.N., Al-Betar, M.A., Abasi, A.K., Awadallah, M.A., Doush, I.A., Alyasseri, Z.A.A. and Alomari, O.A., 2023. Recent advances in butterfly optimization algorithm, its versions and applications. *Archives of Computational Methods in Engineering*, *30*(2), pp.1399-1420.

[14] Kadhm, M.S., Mohammed, M.J. and Zaben, S.O., 2025. Proposed an Accurate Optimization Algorithm Using Butterfly Optimization and Sine-Cosine Optimization Algorithms. *International Journal of Intelligent Engineering & Systems*, *18*(1).

[15] Sadeghian, Z., Akbari, E. and Nematzadeh, H., 2021. A hybrid feature selection method based on information theory and binary butterfly optimization algorithm. *Engineering Applications of Artificial Intelligence*, *97*, p.104079.

[16] Y. Weng, X. Wang, J. Hua, H. Wang, M. Kang and F. -Y. Wang, "Forecasting Horticultural Products Price Using ARIMA Model and Neural Network Based on a Large-Scale Data Set Collected by Web Crawler," in *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 547-553, June 2019

[17] H. Li, Y. Cui, S. Wang, J. Liu, J. Qin and Y. Yang, "Multivariate Financial Time-Series Prediction With Certified Robustness," in *IEEE Access*, vol. 8, pp. 109133-109143, 2020

[18] A. Tiwari, "A Hybrid Feature Selection Method Using an Improved Binary Butterfly Optimization Algorithm and Adaptive β–Hill Climbing," in *IEEE Access*, vol. 11, pp. 93511-93537, 2023

[19] Arora, S. and Anand, P., 2019. Binary butterfly optimization approaches for feature selection. *Expert Systems with Applications*, *116*, pp.147-160.

[20] N. -B. -v. Le, Y. -S. Seo and J. -H. Huh, "AgTech: Volatility Prediction for Agricultural Commodity Exchange Trading Applied Deep Learning," in *IEEE Access*, vol. 12, pp. 153898-153910, 2024,

[21] Yang, X., Zhang, Z. and Xu, H., 2024. RV-FELM: Futures commodity price forecasting based on RIME-VMD algorithm coupled with FA-ELM. *Heliyon*, *10*(17).

[22] Fathollahi-Fard, A.M., Tian, G., Ke, H., Fu, Y. and Wong, K.Y., 2023. Efficient multi-objective metaheuristic algorithm for sustainable harvest planning problem. *Computers & Operations Research*, *158*, p.106304.

[23] J. Liu, B. Zhang, T. Zhang and J. Wang, "Soybean Futures Price Prediction Model Based on EEMD-NAGU," in *IEEE Access*, vol. 11, pp. 99328-99338, 2023

[24] de Amorim, L.B., Cavalcanti, G.D. and Cruz, R.M., 2023. The choice of scaling technique matters for classification performance. *Applied Soft Computing*, Elsevier, *133*, p.109924.

[25] Kurita, T., 2021. Principal component analysis (PCA). In *Computer vision: a reference guide* (pp. 1013-1016). Cham: Springer International Publishing.

[26] Makhadmeh, S.N., Al-Betar, M.A., Doush, I.A., Awadallah, M.A., Kassaymeh, S., Mirjalili, S. and Zitar, R.A., 2023. Recent advances in Grey Wolf Optimizer, its versions and applications. *IEEE Access*.

[27] Nadimi-Shahraki, M.H., Taghian, S. and Mirjalili, S., 2021. An improved grey wolf optimizer for solving engineering problems. *Expert Systems with Applications*, *166*, p.113917.

[28] Liu, Y., As' arry, A., Hassan, M.K., Hairuddin, A.A. and Mohamad, H., 2024. Review of the grey wolf optimization algorithm: variants and applications. *Neural Computing and Applications*, *36*(6), pp.2713-2735.

[29] Makhadmeh, S.N., Al-Betar, M.A., Abasi, A.K., Awadallah, M.A., Doush, I.A., Alyasseri, Z.A.A. and Alomari, O.A., 2023. Recent advances in butterfly optimization algorithm, its versions and applications. *Archives of Computational Methods in Engineering*, *30*(2), pp.1399-1420.

[30] Alweshah, M., Khalaileh, S.A., Gupta, B.B., Almomani, A., Hammouri, A.I. and Al-Betar, M.A., 2022. The monarch butterfly optimization algorithm for solving feature selection problems. *Neural Computing and Applications*, pp.1-15.

[31] Sharma, T.K., 2021. Enhanced butterfly optimization algorithm for reliability optimization problems. *Journal of Ambient Intelligence and Humanized Computing*, *12*(7), pp.7595-7619.