# Hybrid Machine Learning Models For Optimized Potato Price Prediction

Liping Liu[1*]

[1] School of Management, Wuhan College, Wuhan 430212, Hubei, China
[*] Corresponding author. E-mail: llplusida@163.com

Received: Sep. 29, 2024; Accepted: Apr. 04, 2025

Accurate prediction of agricultural commodity prices, such as potatoes, is crucial for enhancing market efficiency, supporting supply chain decisions, and ensuring economic stability in the agricultural sector. This study proposes an enhanced machine learning framework for potato price prediction using Light Gradient Boosting Regression (LGBR), optimized through two metaheuristic algorithms: the Stochastic Paint Optimizer (SPO) and the Population-based Vortex Search Algorithm (PVSA). The hybrid models LGSP (LGBR+SPO) and LGPB (LGBR+PVSA) were developed to reduce prediction error and improve generalization. Experimental results demonstrate that the optimized models outperform the baseline LGBR model. Specifically, LGPB achieved the lowest training mean squared error (MSE) of 3.33E + 03, though it increased to 6.30E + 03 in validation, indicating a potential overfitting issue. LGSP achieved moderate performance with a training MSE of 5.35E+03 and validation MSE of 7.77E+03. In contrast, the baseline LGBR model had the highest MSE values in both training (1.13E+04) and validation (1.34E+04), reflecting weaker predictive accuracy. Uncertainty measures (U95) followed a similar trend. The findings confirm that metaheuristic optimization can significantly improve regression performance in price forecasting tasks. However, challenges in model generalization highlight the need for further tuning and diverse datasets.

**Keywords:** Potato Prices, Decision-Making Process, Machine Learning, Light Gradient Boosting Regression, Stochastic Paint Optimizer.



http://dx.doi.org/10.6180/jase.202605_29(5).0010

## 1. Introduction

For a significant section of the population in the modern developed world, potatoes are a staple diet. It is so widely consumed that, in contrast to its low usage in developing nations, its significance and regular consumption are sometimes overlooked or mistakenly attributed to differences in taste [1]. Using data from recent decades, this article summarizes earlier studies on the subject by showing how, as development advances, the potato goes from being a luxury commodity to a typical good and, ultimately, to a subpar product [2], [3]. It also highlights potato pricing, particularly in underdeveloped regions, as an ideal case for independent study to discover insights that could be applied to topics less accessible through experimental methods due to computational limitations. In the context of agricultural economics, it also offers concepts pertaining to input costs and product prices [4].

The primary source of the potato's spread was Europe, where it originated as a luxury product. However, because it was successfully cultivated and made available to the impoverished, it swiftly moved from being seen as a normal good to becoming a common inferior product throughout Europe [5], [6]. The most popular food among the poor in 19thcentury Europe was definitely potatoes, which acted as a spur for economic expansion and the improvement of welfare generally. Other places also saw the potato's assistance with dietary requirements and city expansion, which sparked development by generating a variety of human capital resources [7], [8]. It is well known that increasing

**Table 1.** Nomenclature

| | | | |
|---|---|---|---|
| Light Gradient Boosting Regression | LGBR | Stochastic Paint Optimizer | SPO |
| Population-based Vortex Search Algorithm | PVSA | Machine Learning | ML |
| mean squared error | MSE | Uncertainty measures | U95 |
| Geographic Information Systems | GIS | Global Positioning Systems | GPS |
| Root Mean Squared Error | RMSE | | |

incomes have led to a more diverse diet and less potato consumption in affluent countries, which is naturally causing ambiguity in people's attitudes toward potatoes today [9], [10]. Many individuals in the developed world consider potatoes to be an essential staple meal and erroneously believe that flavor variations account for the lower usage of potatoes in developing countries. However, the true cause of these disparities in potato use patterns is price. Research environments have also been impacted by historical trends in potato status in western temperate nations, where potatoes flourished.

The potato used to be a major factor in people's social and economic development, particularly for the poor. The Giffen paradox seen in economics textbooks is also oddly reflected in potato pricing, albeit it is unlikely that the potato could be included in a situation where growing costs push the underprivileged to consume more of them. However, the potato is still a widespread food item in many current developing nations, and consumption of it is rising, usually in proportion to wealth, as people's eating habits vary from local staple foods like rice, wheat, or maize to more uncommon and frequently costly potatoes [11], [12].

### 1.1. Related Papers

Since the mid-1920s, the world's population has been growing exponentially. As of October 2018 (www.worldometers.info), there were 7.7 billion people on the planet, and this number is expected to rise by an additional three billion over the next 50 years [13]. The demand for food will increase along with the global population, and competition for arable land and water resources needed to boost agricultural food production is predicted [14]. According to Rijsberman and Molden [15], there is a need to reduce water usage in agriculture by $10-20\%$ while increasing overall food output by about 40%. However, these assumptions must account for the potential consequences of projected climate change, which might significantly impact crop productivity and other critical agricultural resources, like water availability [16]. Land, fossil fuels, and nutrients are other essential resources that ensure food production, even when their present use exceeds the rate of world regeneration. In an attempt to solve significant global concerns, including anthropogenic climate change, the depletion of natural resources, and food security, precision agriculture ( *PA* ) has emerged. Optimizing profits while minimizing the possible environmental effects of farming is the main objective of PA [17].

Reducing the adverse effects of agricultural practices and guaranteeing the future security of food supplies [18] can be achieved by utilizing cutting-edge technologies like satellite data,

Geographic Information Systems (GIS), and Global Positioning Systems (GPS) to enhance crop yield and quality. Satellite remote sensing data is specifically used in agriculture for a number of purposes, such as identifying soil characteristics, classifying crop types, predicting crop yield, tracking crop health, obtaining soil moisture, and assessing meteorological data [19],[20]. Large volumes of information from remote sensing, which is related to big data, can enhance crop modeling and decision-making [21]. "Big data" is defined by Wolfert et al. [22] as large amounts of data with a wide variety that can be acquired, evaluated, and used for decision-making. According to these writers, the agriculture sector is expected to be significantly impacted by big data. Given the volume and diversity of this data, ML has become a useful tool for identifying patterns and rules in datasets [23], as well as for solving non-linear problems on its own [24]. Numerous studies have shown how effective ML techniques are in predicting agricultural production for different varieties, allowing farmers and policymakers to use the best marketing and storage strategies [25], [26], [27]. However, in terms of testing and improving models, tuber and root crops have not gotten much attention up to this point [19].

Over a billion people worldwide eat potatoes, which rank third in importance among food crops after rice and wheat [20]. The rising demand for potatoes and the declining amount of arable land available for growth indicate that improved crop protection and management measures are required to increase crop yields [28]. A variety of management techniques, such as planting dates, population density, irrigation timing and frequency, and fertilizer treatments under various environmental conditions, have long been evaluated for their impact on crop growth and output using crop growth models. Accordingly, crop models might be helpful in improving yield projections for the potato

processing sector [29]. These traditional potato models are commonly used to forecast yields during the growing season and are mostly based on the response to temperature, sunshine, nitrogen fertilizer, and solar radiation incidence. In the literature, there are several models for potato crop growth, such as SUBSTOR-Potato,

LINTUL-Potato, SOLANUM, APSIM Potato, SPUDSIM, POMOD, SIMPOTATO, and Potato Calculator [30], [31], [32]. Some of these models, nevertheless, have never even been used in a real application, and the majority have not been thoroughly evaluated against actual field data [33]. The primary drawbacks of these models are the expense of acquiring the input data needed to run them, the quality of the input data, and sometimes the absence of geographic information [34]. Multispectral satellite imaging in remote sensing can efficiently and across time and space characterize crop development for agricultural production predictions [35]. Consequently, satellites provide a number of ways to lower crop forecasting errors, especially in areas with limited data when input information is unavailable [36], [37].

These models, however, typically require calibration of the local characteristics of the research area. The accuracy levels of previous remote sensing-based potato production models ranged from 0.47 to 0.84 of $R^2$, and they mostly used vegetation indicators from the red and infrared portions of the spectrum. On the other hand, employing different plant indicators based on spectral bands, such as the red-edge ( $\sim 700 - 780\_m$ ), might enhance knowledge of crop status [38]. This spectral region, for instance, has been associated by certain authors with the chlorophyll content or the canopy nitrogen status. Sentinel 2 satellites and other high geographical and temporal resolution satellite images are already freely accessible and have significant promise for crop monitoring and yield forecasting [39].

### 1.2. Research Objectives

Despite significant advances in agricultural price forecasting, predicting potato prices remains a complex task due to the influence of multiple non-linear and volatile factors such as seasonal patterns, market demand, environmental conditions, and policy changes. Existing approaches often rely on traditional or standalone machine learning (ML) models that lack adaptability and robustness when applied to dynamic and uncertain agricultural markets.

Additionally, these models frequently overlook the importance of optimization in improving predictive accuracy and minimizing error. Another notable shortcoming in current studies is the limited use of advanced hybrid modeling techniques that combine strong predictive algorithms with modern metaheuristic optimizers. These gaps hinder the development of highperformance models capable of capturing complex relationships in price data.

To address these challenges, this study proposes a novel hybrid modeling framework that integrates Light Gradient Boosting Regression (LGBR) with two recent optimization algorithms: the Stochastic Paint Optimizer (SPO) and the Population-based Vortex Search Algorithm (PVSA). By hybridizing these components, the study aims to enhance forecasting accuracy and reduce common issues such as overfitting and poor generalization. Furthermore, the research introduces rigorous data preprocessing steps, including the handling of missing values and selection of the most relevant features, to improve model reliability.

The key contribution of this work lies in its ability to combine advanced machine learning techniques with recent optimization strategies, resulting in improved predictive performance. Extensive evaluation using both standard error metrics and uncertainty analysis ensures that the models are tested comprehensively. This approach not only fills the methodological gaps observed in existing literature but also offers practical insights for stakeholders such as farmers, suppliers, and policymakers. Ultimately, the study contributes to the development of more reliable, accurate, and scalable solutions for agricultural price forecasting.

### 1.3. Paper Structure

The structure of this paper is organized to guide readers through the key components of the research. Section 1 introduces the study's background, objectives, and contributions. Section 2 presents the methodology, including model selection, optimization algorithms, and data processing. Section 3 explains the dataset characteristics and preprocessing steps. Section 4 discusses the feature selection techniques and their impact on model performance. Section 5 outlines the evaluation metrics used to assess the models. Section 6 presents and analyzes the experimental results, while Section 7 concludes the study by summarizing key findings, limitations, and potential directions for future work.

## 2. Material and methods

### 2.1. Overview of Study methodology

### 2.2. Study Design

This study was designed to develop and evaluate machine learning models for forecasting potato prices using regression-based approaches. The primary goal was to enhance predictive accuracy through hybridization with metaheuristic optimization algorithms. A comparative experimental design was used, where one baseline model

Light Gradient Boosting Regression (LGBR) was evaluated against two hybrid models: LGBR with Stochastic Paint Optimizer (LGSP) and LGBR with Population-based Vortex Search Algorithm (LGPB). The experiment was structured in two phases: training and validation, using identical data splits to ensure fair comparison.

### 2.3. Implementation Procedure

The study was carried out in the following steps:

- Data Collection & Preprocessing: Historical daily potato price data were obtained from a public agricultural market database. The dataset included date-wise price entries and relevant market features such as season, trend components, and lags. Missing values were handled using forward-fill techniques, and all features were normalized to [ 0,1 ] using MinMax scaling.
- Model Development: Three models were constructed:
- LGBM: The baseline Light Gradient Boosting Machine regression model.
- LGSP: LGBM optimized using the Stochastic Paint Optimizer (SPO) to fine-tune hyperparameters.
- LGPB: LGBM optimized using the Population-based Vortex Search Algorithm (PVSA).

For both SPO and PVSA, the hyperparameters optimized included learning rate, number of estimators, and maximum depth.

- Training & Validation: The dataset was split into 80% for training and 20% for validation. A 5-fold cross-validation scheme was applied during training to reduce overfitting and improve generalizability. Optimization algorithms were applied iteratively during training to minimize the mean squared error (MSE) of the LGBM model.

### 2.4. Data Analysis

Model performance was assessed using the following statistical evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Uncertainty at 95% confidence level (U95) and etc. Each model's training and validation errors were recorded to analyze generalization capability. Comparative analysis helped identify which model achieved the best balance between accuracy and robustness. Further statistical analysis included:

- Trend assessment of performance degradation from training to validation phases.
- Uncertainty quantification using U95 to interpret model stability.
- Overfitting diagnosis by comparing training and validation MSE gaps.

All implementations were conducted using Python 3.10, with the scikit-learn and optuna libraries for ML and optimization, respectively.

### 2.5. Light Gradient Boosting Regression (LGBR)

A recently developed model called LGBM uses boosting and decision trees as a gradient learning framework. In order to minimize memory use, LGBM discretizes continuous floatingpoint eigenvalues into $k$ bins and generates a $k$-width histogram. LGBM keeps accuracy while optimizing the model via a depth-limited leafwise growth approach. Moreover, the method does not need extra storage for results that have already been sorted. Since the segmentation point in a decision tree is a poor learner, its accuracy is irrelevant. In order to minimize overfitting, even greater division scores may have a normalizing effect. The technique is inefficient due to the level-wise development strategy of the decision tree, which handles the leaves of the same layer and consumes much memory that may not be required. The leaf-wise methodology is more effective, according to the level-wise method, since it chooses the foliage from all of the leaves that has the most branching advantage and then continues with the branching process. This approach reduces errors and increases accuracy in an equivalent number of segmentation cycles. The leaf-wise method's drawback is that building deeper decision trees may cause overfitting.

To maintain high efficiency and prevent overfitting [40], a maximum depth limit is included in LGBM at the leaf's top. By analyzing the given training dataset $X = \{(xi, yi)\}i = m1$, LGBM aims to decrease expected values of specified loss coefficients $L(y, f(x))$ and find an approximate $\hat{f}(x)$ of the variable $f^*(x)$.

$$\hat{f}(x) \arg\min_{f} E_{y,x} L(y, f(x)) \tag{1}$$

To evaluate the final model, which is defined as follows, LGBM will make use of multiple regression trees $\sum_{t=1}^{T} f_t(X)$.

$$f_T(x) = \sum_{t=1}^{T} f_t(X) \tag{2}$$

$w_q(x)$, where $q \in \{1, 2, \ldots, N\}$, and $N$ is the number of tree leaves, is how regression trees are represented [41]. The decision rule of the tree is represented by $q$, while the

sample values of the leaf nodes are represented by $w$, a vector. At step $t$, the model is trained additively using the Eq. (3):

$$\eta_t \cong \sum_{j=1}^{N} L\left(y_i, f_{t-1}\left(x_i\right) + f_t\left(x_i\right)\right) \tag{3}$$

The objective function may be rapidly estimated using the Newton technique. Removing the continuous phrase from Eq. (4) reduces it to:

$$\eta_t \cong \sum_{j=1}^{N} \left(u_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right) \tag{4}$$

The mathematical results of the loss functions for the first- and second-order gradients are represented by the variables $u_i$ and $h_i$. Eq. (4) becomes Eq. (5) when the specimen, a collection of leaves $j$, is represented by $I_j$ :

$$\eta_t = \sum_{j=1}^{J} \left(\left(\sum_{i \in I_j} u_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right) \tag{5}$$

The ideal leaf weights, $w_j$, and the maximum values of $\eta_t$ for the tree architecture $q(x)$ are determined by Eq. (6) and Eq. (7):

$$w_j^* = -\frac{\sum_{i \in I_j} u_i}{\sum_{i \in I_j} h_i + \lambda} \tag{6}$$

$$\eta_t^* = -\frac{1}{2} \sum_{j=1}^{J} \frac{\left(\sum_{i \in I_j} u_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} \tag{7}$$

The function that uses weight to assess the structural quality of the tree $q(x)$ is called the weight function. Integrate the split into the subsequent sequence in order to obtain the objective function:

$$O = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_l} u_i\right)^2}{\sum_{i \in I_l} h_i + \lambda} + \frac{\left(\sum_{i \in I_r} u_i\right)^2}{\sum_{i \in I_r} h_i + \lambda} + \frac{\left(\sum_{i \in I} u_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] \tag{8}$$

For the right and left branches, respectively, the symbols $I_r$ and $I_l$ stand for the samples.

### 2.6. Stochastic Paint Optimizer (SPO)

The primary processes of this algorithm are to create the initial paintings, cluster them, combine them, and stop controlling them. Paintings are used as solutions, and the search space is defined as a canvas with certain colors acting as design variables. Based on their respective beauty index (objective function values), the paints are rated and arranged in ascending order. Each additional color added to the canvas becomes an essential component of how the work is perceived. The primary (best), secondary (good), and tertiary (worst) hues on the color wheel correspond to the grades (values) assigned to each color. These equal categories eliminate the need for additional algorithmic factors. The ideal paints (or solutions) can be produced by this algorithm using the given combination strategies to generate new colors.

### 2.7. The Main Steps of the SPO

Phase 1: Setup Random selection determines the starting colors of all paints in an $nc-$ dimensional search object.

$$C_{i,0} = C_{\min} + \text{rand} \cdot \left(C_{\max} - C_{\min}\right), \quad i = 1, 2, 3, \ldots, nc. \tag{9}$$

The starting color of the paint, $C_{i,0}$, is presented while $C_{\min}$ and $C_{\max}$, respectively, indicate the lowest and higher bounds of the design variable. The variables or colors are represented by the variable $nc$, and the random number, rand, has a range of $[0, 1]$. It is important to remember that mixing all the colors together creates paint, which is an optimization issue solution or design. After that, each paint's objective function is evaluated. This brings forth the unique brilliance of each paint.

Phase 2: Evaluation, Sorting, and Clustering: As a result of the task, paints are arranged in ascending order by matching goal functions. Finally, they are classified into three equal categories, namely, primary (the best), secondary (excellent), and tertiary (the poorest), as described in Sect. 2. In this manner, the clustering technique may be used without the need to specify parameters.

Phase 3: Utilizing Combination Techniques: In this stage, four distinct combination approaches are used to develop new paints.

Phase 4: Assessing and revising: If the new paint's beauty index is higher than the old one after evaluation, the new paint will be used in its place.

Phase 5: Verifying termination: The optimization cycle ends when a certain number of iterations are completed. Phase 2 will see the scheduling of a new process if the criterion is not satisfied; if it is, the process will end, and the best option will be presented.

### 2.8. Population-based Vortex Search Algorithm (PVSA)

As a metaheuristic, the VSA is based on a single solution and has effective application characteristics that facilitate rapid implementation [42]. Potential solutions are grouped around a central point by the VSA using a Gaussian distribution to produce new solutions. In some situations, though, this strategy may result in early convergence despite efforts to encourage heterogeneity in the search

area. However, while a search area is still being explored, population-based approaches work best, especially when there are uncharted areas that require further investigation. These methods utilize the data collected from each point in the previous iteration to generate new coordinates [43].

## 3. Initializing

The definition of crucial control parameters occurs during the algorithm's initialization stage. These parameters include the mutation frequency $(\eta_m)$, termination criteria, population size (psize), and vortex size (vsize). The psize parameter represents the total number of candidate solutions generated in a single iteration. It is divided equally in half to obtain vsize, which is equal to psize/2. The vsize value in the first stage is equal to the number of candidate solutions (CS) produced. Next, in the step that follows, more CSs are created, ranging from (vsize +1 ) to psize. When the algorithm reaches the predetermined maximum function evaluations ( max $FEs$ ), it ends. The polynomial mutation process in the second step depends on the probability parameter $\eta_m$. Additionally, Eq. (10) and Eq. (11) are utilized in a certain order to calculate $\mu_0$ and $q_0$ :

$$\mu_0^i = \frac{\text{upper}_{\,i} + \text{lower}_{\,i}}{2} \tag{10}$$

$$q_0^i = \sigma_0^i = \frac{\max(\text{upper}_{\,i}) - \min(\text{lower}_{\,i})}{2} \tag{11}$$

## 4. First phase

According to Eq. (12) of the original VSA, the best solution found is used to update the central point $(\mu)$, where half of the population is generated using a Gaussian distribution. The first iteration of this phase uses random generation to create the entire population of psize, and in subsequent cycles, random generation is limited to vsize, which represents half of the population. After prioritizing the ideal center and applying it to 50% of the population, the other 50% is updated using a population-focused method that incorporates aspects of selection pressure. If a solution exceeds the threshold, it is adjusted to fall within the allowable range using the formula in Eq. (13).

$$s_i^t\left(x_i^t \mid \mu_t, v\right) = \left((2\pi)^d |v|\right)^{-\left(\frac{1}{2}\right)} e^{\left(-1/2\left(x_i^t-\mu_t\right)^T v^{-1}\left(x_i^t-\mu_t\right)\right.} \tag{12}$$

$$s_i\left(\text{lower}_i \vee s_i\right)\text{upper}_i \rightarrow s_i = \text{rand} \times \left(\text{upper}_i - \text{lower}_i\right) + \text{lower}_i \tag{13}$$

Even though it is not a direct component of the original VSA, the first central point, $\mu_0$, helps create the initial population. The candidates chosen from the current population will be the only ones used for the upcoming central point selection. At this point, the VSA is modified, leading to the development of unique PVSA iterations. $PVS_a$ denotes the presence of $\mu_0$ in the original population, whereas $PVS_b$ denotes its absence. In the $PVS_a$ 's first stage, $\mu_0$ takes the place of POP$(1)$, the main candidate solution, in the population, while POP$(2:$ psize$)$, the remaining psize -1 candidate solutions, are generated randomly. In contrast, the starting population for $PVS_b$ is created at random using psize candidate solutions that are chosen from the set $POP$ (1: psize).

## 5. Second phase

Population-based algorithms, in contrast to individual solution algorithms, necessitate interaction between candidate solutions throughout iterations in order to adjust their locations during the search process. The basic method of population-based algorithms is to describe the experiences of individual and collective candidate solutions in vector format, which facilitates information flow. However, the updating process may vary based on the particular algorithm being used. A proportionality-based selection strategy is used in the framework of the PVS algorithm. The update of the position of the candidate solution from the observer bee phase found in the ABC method is combined with particular modifications designed to meet minimization challenges. Eq. (14), for each candidate response, determines the selection probability vector ( $pb$ ).

$$\begin{aligned} pb_i &= \text{csum}_i \,/\, \text{csum}_{psize} \\ \text{csum}_i &= \sum_{j=1}^{i} \text{normp}_i \text{ and} \\ \text{normp}_i &= p_i / \sum_{i=1}^{psize} p_i \text{ and} \\ p_i &= 0.9 \times \left(\max\{\vec{f}\} - f_i\right) + 0.1 \end{aligned} \tag{14}$$

The health metric linked to the solution indexed as " $i$ th" is represented by the symbol $f$. The greatest fitness value in the present population is indicated by $\max\{f^{\rightarrow}\}$. When the values of the objective function are converted from a minimization to a maximization perspective, the modified fitness score of the $i$ th minimization solution, or $p_i$, is determined.

The probabilities that result from modifying the $p$-values so that they lie between 0.5 and 1 are referred to as "Normp." The probabilities obtained by normalizing the

$p$ values are between 0.5 and 1 . A randomly selected close solution is selected from the whole population for the solutions in the latter part of the population, which comprises solutions marked as " $CS_i$ " where " $i$ " is between vsize +1 and psize. Using Eq. (15), the value of a randomly selected dimension is altered to provide a new solution, " $CS_{\text{new}}$ . " Following this modification, as shown in Eq. (16), the resulting dimension value is examined to see whether it exceeds specific standards. This process is guided by the "prob" vector.

$$CS_{\text{new}} = CS_{\text{current}} \text{then} CS^i_{\text{new}} = CS^i_{\text{current}} + \left(CS^i_{\text{current}} - CS^i_{\text{neigbour}}\right) \times (r - 0.5) \times 2 \tag{15}$$

$$CS_{\text{new}} = \begin{cases} \text{lower}_{\,i}, & CS^i_{\text{new}} < \text{lower}_{\,i} \\ \text{CS}^{\,i}_{\text{new}}, & \text{lower}_{\,i} \le CS^i_{\text{new}} \le \text{upper}_{\,i} \\ \text{upper}_{\,i}, & CS^i_{\text{new}} > \text{upper}_{\,i} \end{cases} \tag{16}$$

The fitness of the freshly created solution, " $CS_{\text{new}}$ " is calculated using a random value, r , selected at random from 0.5 to 1 . If the fitness of $CS_{\text{current}}$ , with the former replacing the latter. When the recently computed fitness is compared to the current solution's fitness, "CS current"

Nevertheless, if " $CS_{\text{new}}$ " is unable to exceed " $CS_{\text{current}}$ ," the polynomial mutation method outlined in Eq. (17) is utilized to construct a mutant solution known as "CS mutant."

$$CS_{\text{mutant}} = CS_{\text{current}} + \delta_q \times (\text{ upper } - \text{ lower })$$

$$\delta_q = \begin{cases} \left[ \begin{array}{c} 2r + (1-2r) \\ (1-\delta_1)^{\eta_m+1} \end{array} \right]^{\frac{1}{\eta_m+1}}, & \text{if } r \le 0.5 \\ 1 - \left[ \begin{array}{c} 2(1-r) + 2(r-0.5) \\ (1-\delta_2)^{\eta_m+1} \end{array} \right]^{\frac{1}{\eta_m+1}}, & \text{otherwise} \end{cases} \tag{17}$$

$$\delta_1 = \frac{CS_{\text{current}} - \text{lower}}{\text{upper} - \text{lower}}$$

$$\delta_2 = \frac{\text{upper} - CS_{\text{current}}}{\text{upper} - \text{lower}}$$

When in this case, a random number, known as "rnd," is produced independently for every dimension and ranges from 0.5 to 1 . The next steps are taken if "rnd" turns out to be less than the computed $m$ value, which is obtained by taking the reciprocal of the dimensionality of the issue in question. Prior studies have confirmed that the polynomial mutation operator is the most effective method for resolving the challenging issue of avoiding localized optima and preserving a diversified search space exploration that occurs in metaheuristics. The polynomial mutation operator creates a disturbance effect after employing the polynomial probability distribution to introduce disturbances into the solution. After that, using a selection process that favors the best answer, a comparison between the "CS $_{\text{current}}$ " and "CS mutant" is made. After this process is finished, the center point ( $\mu$ ) is revitalized using the best solution found.

After the current generation is finished, the radius size for the next generation is reduced by using Eq. (18). After reaching the maximum number of function evaluations, the PVS algorithm continues to run. The first step involves duplicating vsize solutions that fall inside the smaller radius, while the second phase introduces stochastic data for the solutions that make up the remaining portion of the population.

$$r_t = \sigma_0 \times \frac{1}{x} \times \Gamma\left(x, a_t\right)$$
$$\text{where } a_t = \frac{(\text{ MaxFEs } - \text{ Fes })}{\text{MaxFEs}} \tag{18}$$

then if $(a_t \le 0)\, a_t = 0.1$

### 5.1. Performance Evaluators

This section describes a variety of measures that may be used to assess the degree of inaccuracy and correlation in hybrid models in order to determine their effectiveness. The MSE, MRAE, and U95 metrics are covered in this article. Below are the formulae for each of these metrics:

$$U95 = \frac{1.96}{n} \sqrt{\sum_{i=1}^{n} (m_i - b_i)^2 + \sum_{j=1}^{n} \left(m_j - b_j\right)^2} \tag{19}$$

$$\text{MRAE} = \frac{1}{n} \sum_{j=1}^{n} \frac{\left|e_j\right|}{\left|A_j - \bar{A}\right|} \tag{20}$$

$$\text{MSE} = \frac{1}{n} \sum_{j=1}^{n} e_j^2 \tag{21}$$

An alternative way to display the variables is as follows: The sample size is represented by $n$, the measured value by $m_i$, and the predicted value by $b_i$.

- The measured value is represented by $\bar{m}$,
- the mean predicted value is represented by $\hbar$.

## 6. Data examination

For the purpose of predicting the price of potatoes, more than 1,000 datasets were really chosen at random from the following link on the Kaggle website: https://www.kaggle.com/datasets/vinayn2/crop-data-of-the-indian-states. These datasets have helped the study's training and testing stages increase the accuracy of the ML models. These data sets have undergone a number of optimization procedures, such as imputation techniques for efficient handling of missing values and min-max scaling to standardize the data. These stages are crucial because they set up the data for a thorough examination and improve model performance by producing more dependable predicted scores.

The features are rescaled using the Min-Max Scaler technique, which typically standardizes the data within a specified range, usually between 0 and 1. This process retains the characteristics with larger scales from the center of the data distribution. Imputation techniques are used to address incomplete data, which could negatively impact model accuracy. These processes-like mean, median, or mode replacement-maintain the dataset's usefulness for ML algorithms by avoiding bias or distortion brought on by missing or unscaled values.

Table 2 provides some of the statistical features of variables involved in the potato price prediction dataset. This table represents the minimum and maximum, average, and standard deviation for input and output variables. Input variables involve a season, state, area in hectares, soil type, pesticide use, pH , temperature, fertilizer use, and rainfall, whereas the output variable is potato price. Each variable's range is displayed by the minimum and maximum. The price ranges from 92.7 to 5,550. The average price is 1,460 , around which the prices deviate considerably, as evidenced by the standard deviation of 856.

The range of the season variable is large, between 1 and 591. This might show temporal variations within this dataset. The state variable, however, represents more geographical representation with an average of about 2.04 entries per state; the soil type is highly heterogeneous in nature, with values ranging from 1 to 613000 . Some of the categorical inputs include pesticide and fertilizer usage, pH levels, and temperature ranges, all of which seem rather limited in their range, hence suggesting controlled uses. Overall, this table underlines a great number of factors driving potato price variations and their statistical characteristics, which are crucial for intuitive level comprehension of data in general and shaping strategies of ML.

The feature selection in Fig. 2 indicates that the Season and Temperature variables alone account for the increases

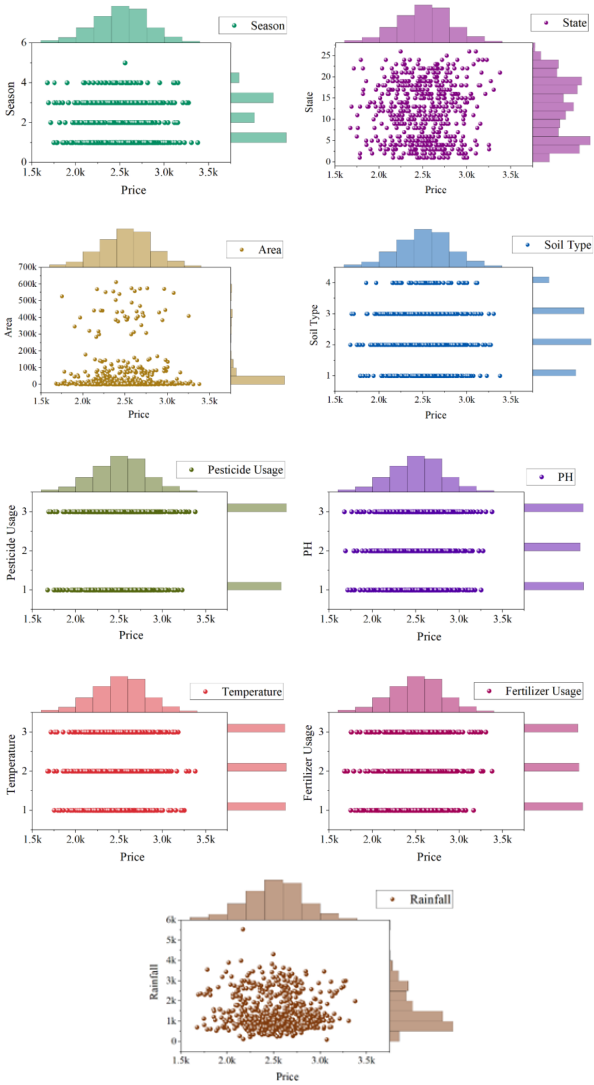


**Fig. 1.** A marginal histogram to show the correlation between the input and outputs.

in variability within the dataset. These two factors introduce greater fluctuations, thus showing that the dataset behaves more erratically to seasonal and temperature variations. This higher variability might be important in understanding the underlying patterns that may point out areas where predictions or estimates are not as reliable since the data is not consistent. In comparison, the soil type, area, and fertilizer usage factors are more consistent and, therefore, less variable. With these variables, there is more consistency in the associated output, which, in turn, shows that their effect on the dataset is more predictable. Whereas Season and Temperature are highly variable, Soil Type, Area, and Fertilizer Usage exhibit stability. This contrast, therefore, suggests that these changes should be put into context in the process of result interpretation or model

**Table 2.** The dataset variables' statistical qualities.

| Variables | Indicators | | | | |
|---|---|---|---|---|---|
| | Category | Min | Max | Avg | St. Dev. |
| Season | Input | 1 | 591 | 296 | $1.71E+02$ |
| State | Input | 1 | 5 | $2.04E+00$ | $1.01E+00$ |
| Area | Input | 1 | 26 | $1.17E+01$ | $6.82E+00$ |
| Soil Type | Input | 3 | $6.13E+05$ | $5.26E+04$ | $1.19E+05$ |
| Pesticide Usage | Input | 1 | 4 | $2.24E+00$ | $9.42E-01$ |
| PH | Input | 1 | 3 | $2.05E+00$ | $9.99E-01$ |
| Temperature | Input | 1 | 3 | $2.00E+00$ | $8.24E-01$ |
| Fertilizer Usage | Input | 1 | 3 | $2.00E+00$ | $8.14E-01$ |
| Rainfall | Input | 1 | 3 | $1.97E+00$ | $8.19E-01$ |
| Price | Output | 92.7 | $5.55E+03$ | $1.46E+03$ | $8.56E+02$ |

design. Indeed, greater attention would likely have to be paid to the variability associated with Season and Temperature in order for predictions to become more reliable, as these factors introduce significant uncertainty into the dataset.

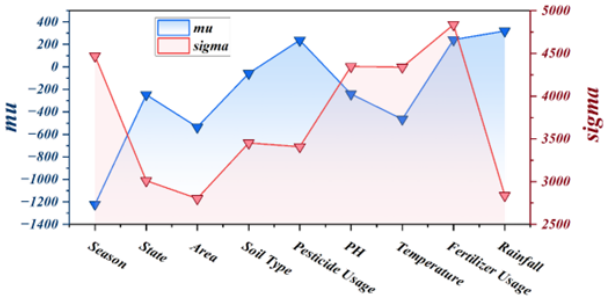


**Fig. 2.** Analysis of feature selection given the input parameters.

### 6.1. Rainfall Impact on Potato Price Prediction

Since rainfall has a direct impact on potato crop yields and quality, it is one of the most important elements impacting price predictions. Potatoes are sensitive to both excessive and insufficient rainfall, as the optimal growth of tubers requires just the right amount of water. Heavy rainfall can lead to soil waterlogging, increasing the risk of fungal diseases and reducing the overall quality of the yield. On the other hand, a lack of rainfall can cause drought stress, leading to smaller tubers and lower yields. These fluctuations in crop production significantly affect the supply chain, which, in turn, impacts market availability and pricing. Consequently, any ML model for potato price prediction must include rainfall data as a crucial feature to capture its influence on the volatility of supply. Additionally, changes in regional rainfall patterns can result in localized price variations, making rainfall a key component of a price prediction model in agricultural economics. Ultimately, the integration of meteorological data, including rainfall patterns, can greatly enhance the accuracy of potato price forecasting [44].

## 7. Results and analysis

To determine which model was the best-performing and most functional, a comprehensive comparison was conducted across various phases of development, including training, validation, testing, and All. This multi-faceted approach ensured a thorough evaluation of each model's effectiveness. Furthermore, a range of quantitative metrics were employed to assess the models' performance, such as Mean Squared Error (MSE), Upper 95 Percent Confidence Interval (U95), and Mean Relative Absolute Error (MRAE). These metrics provide insightful information on the accuracy and dependability of the models, enabling a thorough evaluation of their capabilities and assisting in the selection of the best model for real-world use.

Fig. 3 shows the convergence of the two models (LGSP and LGPB) over 200 iterations. This time, the focus will be on the RMSE values. The vertical axis shows the convergence in terms of RMSE values, while the horizontal axis represents the iteration number. The result highlights that both LGSP and LGPB have a decreasing trend in their RMSE with iteration. However, they start at 7.97E + 01 and 6.69E + 01 for LGSP and LGPB , respectively. LGPB has continuously performed better than LGSP through all the iterations, which resulted in faster convergence at lower RMSE. By the end of 200 iterations, the performance of LGPB significantly outperformed that of LGSP by stabilizing with a lower RMSE. That would mean LGPB is better and more efficient in minimizing error during the optimization process, showing that it might turn out to be the more reliable model when predictive accuracy comes into play for this context.

Fig. 4 illustrates how the measured and predicted values of the generated models correlate: Both the LGSP and

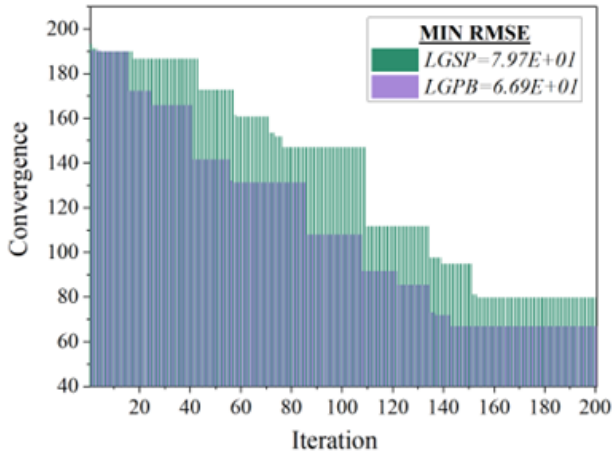


**Fig. 3.** Convergence of hybrid models using column plot

LGPDB models perform well overall, as seen by the substantial positive correlations between the predicted and observed values. While LGSP presents a close clustering around the center line in its prediction, the LGPDB prediction represents fairly similar performance but indicates more accurate predictions in the lower range of the predicted value. The RMSE for both is low; however, that for LGPDB is much lower, indicating higher accuracy. Meanwhile, $R^2$ values are higher for LGPDB, which explains more variation in data compared to LGSP.

On consistency, both models perform well on the training and validation sets, while LGPDB performs better in generalization, especially at an unseen test, keeping a better fit. Comparing the two scatter plots, the points from the LGPDB model lie closer to the center line, especially for some points where the predictions from LGSP are slightly scattered. That would be indicative of how the LGPDB model handles certain ranges of data with superior performance. However, from everything mentioned, LGPDB works better because of the improved accuracy, the higher explained variance, and the better fitting; hence, it is more reliable to use in making a prediction.

A comparison of the LGMB model with previously assessed models, namely LGSP and LGPDB, shows that on critical counts, LGMB scores are higher than those of them. Overall performance is good for all three models, as seen by the substantial positive correlation between predicted and measured values. However, as shown, the RMSE of LGMB is significantly smaller compared to LGSP and LGPDB, reflecting the superior prediction accuracy provided by LGMB. Besides this fact, the $R^2$ values for LGMB are comparably high, meaning it explains a substantial proportion of the variance in the data, similar to that from the others but with a pronounced edge.

By the scatter plot comparison, the data points of LGMB are closer to the center line than the data points of LGSP and LGPDB, hence confirming that it is an extremely good fit for the model. While LGPDB was somewhat better than LGSP on account of its lower RMSE and better treatment of low-value predictions, it is grossly outclassed by LGMB in accuracy and fit. After all, the LGMB model turns out to be the best because it has the lowest RMSE and the highest $R^2$, along with tighter clustering of the predictions. This provides the most accurate prediction that generalizes the best across datasets.

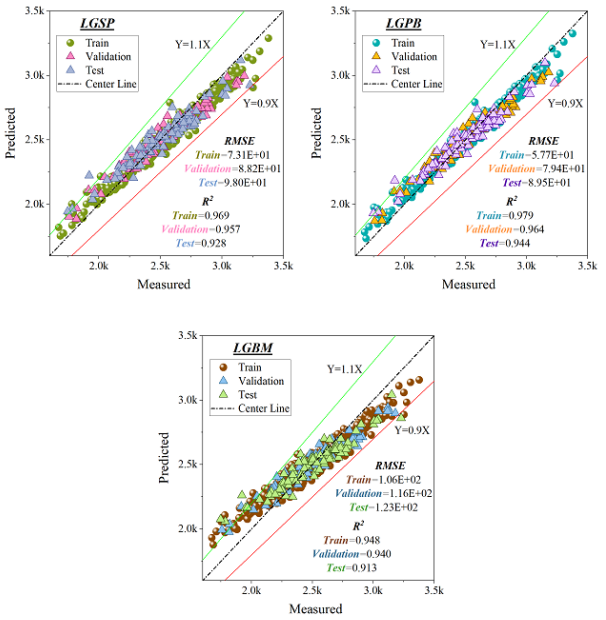


**Fig. 4.** The scatter plot shows how the measured and predicted values are correlated

Table 3 shows the comparative analysis of the models LGBM adds to its MSE from training-1.13E+04 to the validation phase at 1.34E+04. It can thus be inferred that this model has been subjected to a slight decline in generalization performance. It also increases its U95 value from $2.94E+02$ to $3.21E+02$, which indeed points toward higher uncertainty in the predictions during validation.

On the other hand, the most obvious generalization from training to validation is for LGPB, where MSE increases from 3.33E+03 to 6.30E+03, while the U95 increases from 1.60E+02 to $2.20E+02$, reflecting heightened uncertainty in the validation predictions. Meanwhile, LGSP is also showing an increasing trend with the MSE from $5.35E+03$ for training to $7.77E+03$ in validation, which may indicate possible overfitting or poor generalization.

The value of U95 further increases from $1.99E+02$ to $2.44E+02$, showing higher uncertainty during validation.

Overall, all the models show an increase in MSE and U95 from training to validation, indicating difficulties with generalization. LGBM exhibits the highest MSE in both phases, while LGPB demonstrates the smallest MSE during training but experiences a sharp increase in validation, suggesting overfitting. LGSP shows average MSE values, but similar generalization issues persist. Further tuning of model parameters and validation on more diverse datasets may help improve the performance of all these models on unseen data.

Performances in the training, validation, and test sets are shown for comparisons of error distributions of models LGSP, LGPDB, and LGMB in Fig. 5. In general, all three models demonstrate roughly bell-shaped distributions of errors, indicating normal-like behavior, which is desirable for any prediction. Each model maintains peaks at zero, reflecting central tendency and showing overall predictive accuracy. When comparing the dispersion of error distributions across the three models, it is clear that the validation and test sets have a larger spread compared to the training set, which is a common characteristic, as models tend to perform slightly worse on unseen data. Nevertheless, out of the three, the LGMB model performs better on the test and validation sets than on the training data because it has the narrowest dispersion. This is further supported by its fewer outliers in the test set, indicating higher robustness to extreme values compared to the LGSP and LGPDB models. In contrast, the LGPDB model exhibits more outliers than the LGMB, suggesting a slight vulnerability to atypical data points. The LGSP model shows the widest spread in terms of errors and outliers, which may indicate more issues with prediction accuracy, especially in the testing phase.

The performance of the three models in this comparison is shown in Fig. 6 through the RMSE, the MARE, and MSE in the LGSP, LGPDB, and LGMB models. Both metrics have been used to show that the LGSP has the highest error value of the three; hence, it is generally less accurate. This is further confirmed by its corresponding MSE values, which are also the highest and indicate larger squared errors. The LGPDB model outperformed the LGSP model because its values for RMSE and MRAE were relatively lower. This will mean it has some predictive inexactitudes, but its relative errors and squared errors are reduced compared to the LGSP model. Whereas the LGMB model has outperformed for all metrics analyzed, these two models have always been outperformed by the mentioned LGMB model. Having the minimum value for RMSE, MRAE, and MSE, the LGMB model can give more accurate predictions with minimum error, making this model highly reliable compared to the other two models, LGSP and LGPDB. The findings show that while the LGSP and LGPDB models produce more mistakes, the LGMB model has the greatest prediction accuracy. Specifically, the LGSP model urgently needs to be refined since it tends to be less reliable compared with its counterparts. The overall performance shows that LGMB is quite effective in giving high-accurate predictions with the least values from the relative errors and squared errors, becoming the most preferred choice in context.

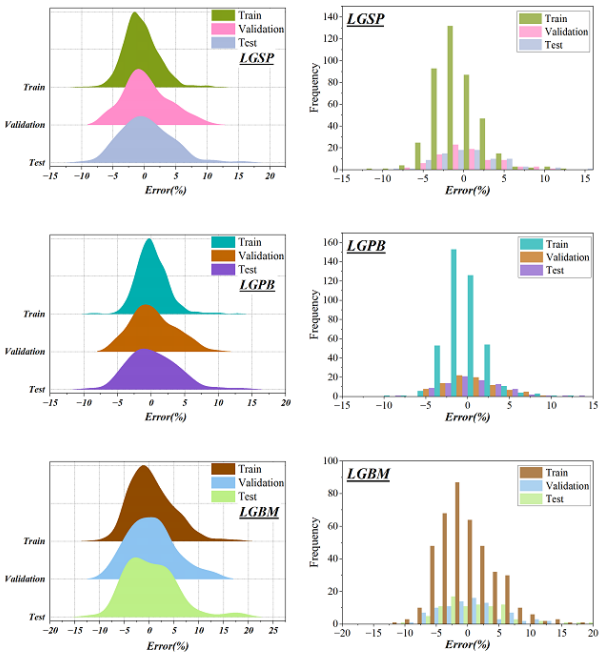


**Fig. 5.** The Distribution and Ridgeline plots are used to determine the error percentage of the models that are displayed

### 7.1. Limitation of Study

While the proposed hybrid models (LGSP and LGPB) demonstrated improved prediction accuracy over the baseline LGBR model, several limitations should be acknowledged:

- Generalization Issues: All models exhibited an increase in mean squared error (MSE) and uncertainty (U95) from training to validation, indicating challenges in generalizing to unseen data. In particular, the LGPB model showed strong performance during training but experienced a significant increase in error during validation, suggesting a risk of overfitting.
- Data Dependency: The model performance is highly

**Table 3.** Comparison of time performance and performance with different superpixel methods

| Model name | Modeling Phase | Index values | | |
|---|---|---|---|---|
| | | MSE | U95 | MRAE |
| LGBM | Train | $1.13E+04$ | $2.94E+02$ | 1.025 |
| | Validation | $1.34E+04$ | $3.21E+02$ | 1.023 |
| | Test | $1.52E+04$ | $3.41E+02$ | 11.270 |
| | All | $1.22E+04$ | $3.06E+02$ | 1.094 |
| LGPB | Train | $3.33E+03$ | $1.60E+02$ | 0.670 |
| | Validation | $6.30E+03$ | $2.20E+02$ | 0.605 |
| | Test | $8.00E+03$ | $2.48E+02$ | 4.282 |
| | All | $4.47E+03$ | $1.85E+02$ | 0.732 |
| LGSP | Train | $5.35E+03$ | $1.99E+02$ | 0.692 |
| | Validation | $7.77E+03$ | $2.44E+02$ | 0.717 |
| | Test | $9.60E+03$ | $2.72E+02$ | 2.562 |
| | All | $6.35E+03$ | $2.19E+02$ | 0.700 |

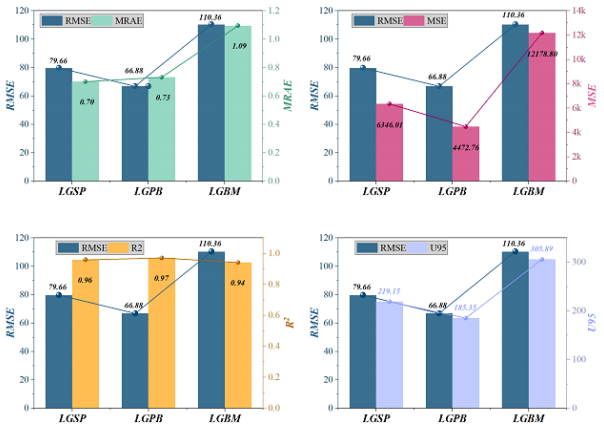


**Fig. 6.** Double-column graphic to evaluate the models' performance

dependent on the quality, size, and diversity of the training dataset. Limited variability in the dataset may constrain the model's ability to capture broader market dynamics or seasonal patterns in potato pricing.

- Computational Complexity: The use of metaheuristic optimization algorithms like SPO and PVSA increases the computational cost and training time. These models may be less practical for real-time applications unless computational resources are adequately scaled.

- Lack of External Validation: The current evaluation was performed on a single dataset. The absence of external validation on datasets from different regions or timeframes limits the model's ability to generalize across markets.

### 7.2. Implications of the findings

The results of this study have several practical and theoretical implications:

- Enhanced Price Forecasting for Agricultural Planning: The integration of machine learning with metaheuristic optimization (LGBR + SPO/PVSA) demonstrates a powerful approach for improving the accuracy of agricultural price predictions. This can directly benefit farmers, traders, and government agencies by supporting better planning for planting, harvesting, storage, and distribution.

- Policy and Market Stability: Accurate price prediction can contribute to more stable agricultural markets. For policymakers, these insights can guide decisions around subsidies, import/export regulations, and emergency food planning to ensure economic resilience and food security.

- Methodological Advancement: The hybridization of LGBR with SPO and PVSA illustrates a novel methodological contribution, showing how metaheuristic algorithms can enhance standard regression techniques in capturing nonlinear trends in agricultural data.

- Scalability to Other Crops and Markets: Although this study focuses on potato prices, the proposed framework can be generalized to other crops or commodities. This opens pathways for broader adoption of machine learning in agroeconomic forecasting tasks.

- Platform for Future Research: The findings provide a baseline for future research into ensemble learning, feature selection, and hyperparameter optimization using different algorithms or cross-regional datasets. Moreover, improving model interpretability could enhance trust and adoption by non-technical stakeholders.

## 8. Concluding insights

The application of machine learning (ML) algorithms to forecast potato prices has proven effective in supporting agricultural decision-making, improving supply chain operations, and enhancing market stability. This study proposed a predictive framework based on Light Gradient Boosting Regression (LGBR), further enhanced using two metaheuristic optimization algorithms: the Stochastic Paint Optimizer (SPO) and the Population-based Vortex Search Algorithm (PVSA). The hybrid models LGSP and LGPB demonstrated superior performance over the standalone LGBM model. Among the evaluated models, LGPB achieved the best performance with an RMSE of 66.88, followed by LGSP with 79.66, and LGBM with 110.36, indicating the effectiveness of optimization in boosting predictive accuracy. Detailed experimental analysis revealed that while all models experienced a performance drop from training to validation, LGBM exhibited the highest error and uncertainty across phases. LGPB, despite the lowest training error, showed signs of overfitting during validation, highlighting a need for improved generalization. LGSP offered balanced performance but still faced increased uncertainty in unseen data. These findings highlight three key contributions: (1) the benefit of integrating metaheuristic optimizers with LGBR for agricultural forecasting; (2) the importance of selecting appropriate optimization techniques based on validation performance; and (3) the need for robust evaluation to ensure model generalizability. However, limitations were also observed. The proposed models depend heavily on the availability of large, highquality datasets. Furthermore, their "black box" nature makes them difficult to interpret for non-technical stakeholders. Additionally, external factors like policy shifts, weather anomalies, or sudden market disruptions are not explicitly modeled, which may limit predictive reliability. Finally, model performance may degrade over time, requiring frequent retraining with updated data, which can pose resource challenges.

## 9. Future work suggestion

Future research should focus on the following directions:

- Integration of External Variables: Incorporating climate data, global market indices, and policy indicators to capture the influence of non-linear external shocks.
- Explainability and Transparency: Employing model interpretability tools (e.g., SHAP values, LIME) to provide insights into feature importance and improve stakeholder trust.
- Real-time Forecasting Frameworks: Developing adaptive models capable of online learning to dynamically respond to changing market conditions.
- Cross-Commodity Forecasting: Expanding the model framework to predict prices of other essential crops, allowing a broader application in agricultural economics.

## 10. Funding

This work was supported by 2024 philosophy and social science project of Hubei Provincial Department of Education (Number:24G108) (Empirical Study on the Factors Influencing the Willingness to Adopt Artificial Intelligence Generated Content (AIGC): A Revision Based on the UTAUT Model).

## References

[1] J. E. da Silva Ribeiro, A. G. C. da Silva, J. V. L. Lima, P. H. de Almeida Oliveira, E. dos Santos Coêlho, L. M. da Silveira, and A. P. B. Júnior, (2024) *"Leaf area prediction of sweet potato cultivars: An approach to a non-destructive and accurate method"* **South African Journal of Botany *172***: 42–51. DOI: https://doi.org/10.1016/j.sajb.2024.07.006.

[2] X. Lei, X. Xu, and S. Zhou, (2025) *"Potato Yield Prediction Research Based on Improved Artificial Neural Networks Using Whale Optimization Algorithm"* **Potato Research *68***: 1717–1726. DOI: https://doi.org/10.1007/s11540-024-09819-9.

[3] M. Hassan, K. Khosravi, A. A. Farooque, T. J. Esau, A. Boluwade, and R. Sadiq, (2024) *"Prediction of carbon dioxide emissions from Atlantic Canadian potato fields using advanced hybridized machine learning algorithms–Nexus of field data and modelling"* **Smart Agricultural Technology *9***: 100559. DOI: https://doi.org/10.1016/j.atech.2024.100559.

[4] P. Chaukhande, S. K. Luthra, R. N. Patel, S. R. Padhi, P. Mankar, M. Mangal, J. K. Ranjan, A. U. Solanke, G. P. Mishra, and D. C. Mishra, (2024) *"Development and validation of near-infrared reflectance spectroscopy prediction modeling for the rapid estimation of biochemical traits in potato"* **Foods *13***: 1655. DOI: https://doi.org/10.3390/foods13111655.

[5] Y. Wang, Y. Xu, X. Wang, H. Wang, S. Liu, S. Chen, and M. Li, (2024) *"Optimizing the effects of potato size and shape on near-infrared prediction models of potato quality using a linear-nonlinear algorithm"* **Journal of Food Composition and Analysis *135***: 106679. DOI: https://doi.org/10.1016/j.jfca.2024.106679.

[6] P. Jha, D. Dembla, and W. Dubey, (2024) *"Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model"* **Multimedia Tools and Applications *83***: 37839–37858. DOI: https://doi.org/10.1007/s11042-023-16993-4.

[7] E.-S. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, (2025) *"Predicting potato crop yield with machine learning and deep learning for sustainable agriculture"* **Potato Research *68***: 759–792. DOI: https://doi.org/10.1007/s11540-024-09753-w.

[8] S. A. Alzakari, A. A. Alhussan, A.-S. T. Qenawy, A. M. Elshewey, and M. Eed, (2025) *"An enhanced long short-term memory recurrent neural network deep learning model for potato price prediction"* **Potato Research *68***: 621–639. DOI: https://doi.org/10.1007/s11540-024-09744-x.

[9] A. Gupta, A. Chug, and A. P. Singh, (2024) *"Potato disease prediction using machine learning, image processing and IoT–a systematic literature survey"* **Journal of Crop Improvement *38***: 95–137. DOI: https://doi.org/10.1080/15427528.2023.2285827.

[10] A. A. Abdelhamid, A. A. Alhussan, A.-S. T. Qenawy, A. M. Osman, A. M. Elshewey, and M. Eed, (2024) *"Potato harvesting prediction using an Improved ResNet-59 model"* **Potato Research**: 1–20. DOI: https://doi.org/10.1007/s11540-024-09773-6.

[11] M. Piekutowska and G. Niedbała, (2025) *"Review of methods and models for potato yield prediction"* **Agriculture *15***: 367. DOI: https://doi.org/10.3390/agriculture15040367.

[12] A. Mukiibi, A. T. B. Machakaire, A. C. Franke, and J. M. Steyn, (2025) *"A systematic review of vegetation indices for potato growth monitoring and tuber yield prediction from remote sensing"* **Potato research *68***: 409–448. DOI: https://doi.org/10.1007/s11540-024-09748-7.

[13] S. K. Seelan, S. Laguette, G. M. Casady, and G. A. Seielstad, (2003) *"Remote sensing applications for precision agriculture: A learning community approach"* **Remote sensing of environment *88***: 157–169. DOI: https://doi.org/10.1016/j.rse.2003.04.007.

[14] H. Lotze-Campen, C. Müller, A. Bondeau, S. Rost, A. Popp, and W. Lucht, (2008) *"Global food demand, productivity growth, and the scarcity of land and water resources: a spatially explicit mathematical programming approach"* **Agricultural Economics *39***: 325–338. DOI: https://doi.org/10.1111/j.1574-0862.2008.00336.x.

[15] R. FR. "Balancing water uses: water for food and water for nature. Thematic background paper". In: *International Conference on Freshwater. December 2001, Bonn, Germany*. 2001.

[16] G. C. Nelson, H. Valin, R. D. Sands, P. Havlík, H. Ahammad, D. Deryng, J. Elliott, S. Fujimori, T. Hasegawa, and E. Heyhoe, (2014) *"Climate change effects on agriculture: Economic responses to biophysical shocks"* **Proceedings of the National Academy of sciences *111***: 3274–3279. DOI: https://doi.org/10.1073/pnas.1222465110.

[17] P. Loudjani, (2014) *"Precision Agriculture: An Opportunity for EU-Farmers–Potential Support with the CAP 2014-2020"*: 1–50.

[18] R. Gebbers and V. I. Adamchuk, (2010) *"Precision agriculture and food security"* **Science *327***: 828–831. DOI: https://doi.org/10.1126/science.1183899.

[19] R. Raymundo, S. Asseng, R. Robertson, A. Petsakos, G. Hoogenboom, R. Quiroz, G. Hareau, and J. Wolf, (2018) *"Climate change impact on global potato production"* **European Journal of Agronomy *100***: 87–98. DOI: https://doi.org/10.1016/j.eja.2017.11.008.

[20] A. Devaux, P. Kromann, and O. Ortiz, (2014) *"Potatoes for sustainable global food security"* **Potato research *57***: 185–199. DOI: https://doi.org/10.1007/s11540-014-9265-1.

[21] L. Sharma, S. K. Bali, and J. D. Dwyer, (2017) *"Study of Improving Yield Prediction and Sulfur Deficiency Detection Using Optical Sensors"* **ASA, CSSA and SSSA International Annual (2017)**: DOI: http://dx.doi.org/10.3390/s17051095.

[22] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, (2017) *"Big data in smart farming–a review"* **Agricultural systems *153***: 69–80. DOI: https://doi.org/10.1016/j.agsy.2017.01.023.

[23] D. Zhang and J. J. P. Tsai. *Advances in machine learning applications in software engineering*. Igi Global, 2006. DOI: http://dx.doi.org/10.4018/978-1-59140-941-1.

[24] A. Chlingaryan, S. Sukkarieh, and B. Whelan, (2018) *"Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review"* **Computers and electronics in agriculture *151***: 61–69. DOI: https://doi.org/10.1016/j.compag.2018.05.012.

[25] S. S. Dahikar and S. V. Rode, (2014) *"Agricultural crop yield prediction using artificial neural network approach"* **International journal of innovative research in electrical, electronics, instrumentation and control engineering 2**: 683–686.

[26] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, (2016) *"Wheat yield prediction using machine learning and advanced sensing techniques"* **Computers and electronics in agriculture *121***: 57–65. DOI: https://doi.org/10.1016/j.compag.2015.11.018.

[27] S. Veenadhari, B. Misra, and C. D. Singh. "Machine learning approach for forecasting crop yield based on climatic parameters". In: *2014 International Conference on Computer Communication and Informatics*. IEEE, 2014, 1–5. DOI: https://doi.org/10.1109/ICCCI.2014.6921718.

[28] W. Bowen, H. Cabrera, V. H. Barrera, and G. Baigorria, (1999) *"Simulating the response of potato to applied nitrogen"*:

[29] A. T. B. Machakaire, J. M. Steyn, D. O. Caldiz, and A. J. Haverkort, (2016) *"Forecasting yield and tuber size of processing potatoes in South Africa using the LINTUL-potato-DSS model"* **Potato research *59***: 195–206. DOI: https://doi.org/10.1007/s11540-016-9321-0.

[30] S. A. Yadav, X. Zhang, N. K. Wijewardane, M. Feldman, R. Qin, Y. Huang, S. Samiappan, W. Young, and F. G. Tapia, (2025) *"Context-Aware Deep Learning Model for Yield Prediction in Potato Using Time-Series UAS Multispectral Data"* **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**: DOI: https://doi.org/10.1109/JSTARS.2025.3539217.

[31] M. Eed, A. A. Alhussan, A.-S. T. Qenawy, A. M. Osman, A. M. Elshewey, and R. Arnous, (2025) *"Potato consumption forecasting based on a hybrid stacked deep learning model"* **Potato Research *68***: 809–833. DOI: https://doi.org/10.1007/s11540-024-09764-7.

[32] R.-F. Wang and W.-H. Su, (2024) *"The application of deep learning in the whole potato production Chain: A Comprehensive review"* **Agriculture *14***: 1225. DOI: https://doi.org/10.3390/agriculture14081225.

[33] D. Borus, D. Parsons, M. Boersma, H. Brown, and C. Mohammed, (2018) *"Improving the prediction of potato productivity: APSIM-Potato model parameterization and evaluation in Tasmania, Australia"* **Australian Journal of Crop Science *12***: 32–43.

[34] R. Radha and M. K. Singh, (2023) *"Axial groundwater contaminant dispersion modeling for a finite heterogeneous porous medium"* **Water *15***: 2676. DOI: https://doi.org/10.3390/w15142676.

[35] M. M. Awad, (2019) *"Toward precision in crop yield estimation using remote sensing and optimization techniques"* **Agriculture *9***: 54. DOI: https://doi.org/10.3390/agriculture9030054.

[36] V. UmaRani and S. Thirisaa. "Analysis of Pre-Trained CNN Models for Pepper and Potato Leaf Disease Prediction". In: *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, 2024, 1–5. DOI: https://doi.org/10.1109/ESCI59607.2024.10497250.

[37] K. Tatsumi and T. Usami, (2024) *"Plant-level prediction of potato yield using machine learning and unmanned aerial vehicle (UAV) multispectral imagery"* **Discover Applied Sciences *6***: 649. DOI: https://doi.org/10.1007/s42452-024-06362-7.

[38] J. G. P. W. Clevers and A. A. Gitelson, (2013) *"Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and-3"* **International Journal of Applied Earth Observation and Geoinformation *23***: 344–351. DOI: https://doi.org/10.1016/j.jag.2012.10.008.

[39] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, (2020) *"Sentinel-2 data for land cover/use mapping: A review"* **Remote sensing *12***: 2291. DOI: https://doi.org/10.3390/rs12142291.

[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, (2017) *"Lightgbm: A highly efficient gradient boosting decision tree"* **Advances in neural information processing systems *30***:

[41] X. Sun, M. Liu, and Z. Sima, (2020) *"A novel cryptocurrency price trend forecasting model based on LightGBM"* **Finance Research Letters *32***: 101084. DOI: https://doi.org/10.1016/j.frl.2018.12.032.

[42] B. Doğan and T. Ölmez, (2015) *"A new metaheuristic for numerical function optimization: Vortex Search algorithm"* **Information sciences *293***: 125–145. DOI: https://doi.org/10.1016/j.ins.2014.08.053.

**[43]** T. Sağ, (2022) *"PVS: a new population-based vortex search algorithm with boosted exploration capability using polynomial mutation"* **Neural Computing and Applications *34***: 18211–18287. DOI: https://doi.org/10.1007/s00521-022-07671-x.

**[44]** S. S. Adudotla, P. Bobba, Z. Pathan, T. Kata, C. C. Sobin, and Jahfar. "A Method for Price Prediction of Potato Using Deep Learning Techniques". In: *International Conference on Intelligent Vision and Computing*. Springer, 2022, 619–629. DOI: https://doi.org/10.1007/978-3-031-31164-2_53.