Discover Computing

RESEARCH Open Access

Explainable AI for economic performance prediction evidence from industrial production, institutional quality, and consumer demand

Oluwaseun Racheal Ojekemi[2*] and Dervis Kirikkaleli[1,2]

*Correspondence:
Oluwaseun Racheal Ojekemi
22246145@std.eul.edu.tr
[1]Adnan Kassar School of Business, Lebanese American University, Beirut, Lebanon
[2]Faculty of Economics and Administrative Science, European University of Lefke, Mersin 10, Lefke, Northern Cyprus, Turkey

Abstract
The GWO-XGBoost model achieved an R^2 of 0.991 in Gross Domestic Product (GDP) prediction, exceeding all other machine learning models compared in this study. Thus, GWO-XGBoost provides a transparent and reliable decision-making support system for all those who develop economic policies. Developing accurate macroeconomic predictions can assist in the advancement of sustainable development. However, the non-linear relationships among many macroeconomic variables, such as industrial production, governance, and demand, make it difficult to develop accurate and understandable predictions using conventional macroeconomic models. To address the limitations of these traditional models, we propose a hybrid explanatory model, GWO-XGBoost, that integrates the Grey Wolf Optimizer (GWO) with Extreme Gradient Boosting (XGBoost). This integrated model automatically tunes its hyperparameters. Monthly US data from 1996 to 2020 was utilized in the experiment. These data include: Government Effectiveness, Consumer Demand/Retail Sales, Industrial Production, Trade Balance, and Unemployment. The results of the SHapley Additive exPlanations (SHAP) analysis show that the two most significant factors influencing economic performance were consumer demand and government effectiveness. As a result of being both highly predictive and having clear feature attributions, the proposed hybrid explanatory model will be useful for all institutions involved in developing and implementing evidence-based planning processes that align with economic growth.

Keywords Explainable AI, GWO–XGBoost, SHAP, GDP prediction, Institutional quality, Consumer demand

1 Introduction

Forecasts of national economic output guide fiscal policy, monetary decisions, and private investment. Accurate gross domestic product (GDP) projections allow governments to set budgetary priorities, adjust interest rates, and deploy counter-cyclical measures during downturns [1]. They also allow firms to plan capital expenditures and employment levels. A recent assessment of machine learning for GDP forecasting notes that

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

precise GDP projections can help governments and businesses allocate resources and implement timely interventions, and that these forecasts are vital [2].

Traditional econometric models, such as the seasonal autoregressive integrated moving average (SARIMA) model, assume linearity and stationarity in the underlying processes [2, 3]. In complex economies, these assumptions often fail, limiting the usefulness of linear models when sudden shocks or structural shifts occur [2, 4, 5]. The rise of big data and real-time analytics has expanded the sources of information available to forecasters. New data streams, including real-time financial transactions, provide granular signals that can improve predictions. Choi and Varian's work demonstrated that this data could augment traditional economic indicators and improve GDP prediction [6]. Economic researchers have therefore begun to explore machine learning (ML) methods capable of capturing nonlinear patterns and high-dimensional interactions [2]. ML has transformed numerous domains by learning intricate patterns within the data and adjusting to varying conditions within the data.

In fields as diverse as healthcare, communications, agriculture, and transportation, ML has enabled better decision-making and optimized resource use [2]. In economics, ML has been applied to forecast variables such as GDP and inflation, offering more flexible representations than conventional econometric models [2, 7]. Long short-term memory (LSTM) networks, a class of recurrent neural networks, have been shown to outperform SARIMA on GDP forecasting tasks in both accuracy and ability to capture long-term dependencies [2, 8, 9]. Other ML techniques, such as support vector machines (SVMs), Random Forests (RF), and Gradient Boosting (GB), have yielded performance gains over linear models in several studies [10]. However, adopting ML in macroeconomic contexts introduces challenges. Models such as Extreme Gradient Boosting (XGBoost) offer high predictive accuracy and robustness to missing data and overfitting [11] but contain many hyperparameters that influence their behavior. Finding appropriate hyperparameter settings is laborious, and poor tuning can negate potential gains. Moreover, many ML models behave as black boxes, making it difficult for policymakers to trust their outputs or understand which factors drive the forecasts.

Accurate forecasting also depends on recognizing institutional factors that shape economic performance. Governance quality is one such factor. The World Bank's Worldwide Governance Indicators (WGI) cover six dimensions of governance for over 200 economies and emphasize that good governance fosters economic growth, builds human capital, and strengthens social cohesion [12]. Among these dimensions, government effectiveness captures perceptions of the quality of public services, the competence and independence of the civil service, the quality of policy formulation and implementation, and the credibility of the government's commitment to those policies. Research indicates that institutions and governance affect long-term growth by shaping incentives and promoting efficient divisions of labor. Efficient governments enhance market efficiency by enforcing property rights and complementing private-sector activities by directing resources and facilitating learning. Empirical evidence on the impact of government effectiveness on growth has been mixed, but recent analyses using dynamic panel methods suggest that government effectiveness has a substantial positive effect on economic growth [13]. A study sub-Saharan African countries found that higher government effectiveness contributed positively and significantly to economic growth [14]. Despite its potential importance, governance quality rarely appears as a feature in macroeconomic

forecasting models, leaving a gap in the integration of institutional variables with data-driven forecasting methods.

The literature on macroeconomic forecasting has evolved from linear time-series models to high-dimensional, data-driven techniques. Chi et al. emphasized three ideas: exploiting large datasets with regularization, rigorously evaluating models out-of-sample, and allowing for nonlinearities [15]. Yet even with large datasets, linear models remain constrained by their functional forms. When economic relationships are nonlinear or vary over time, linear models can underperform. Gradient boosting algorithms such as XGBOOST have become particularly attractive for economic forecasting. XGBOOST combines many weak learners (decision trees) in a boosting framework, employs regularization to control overfitting, and handles missing values internally [11]. However, its performance depends on hyperparameters controlling tree depth, learning rates, and subsampling ratios. Selecting these parameters manually or via grid search is expensive and may fail to explore the solution space adequately. Studies have therefore turned to metaheuristic optimization. Hyperparameter tuning is essential for achieving high performance in ML models. Traditional grid search or random search methods can be inefficient because they explore only a small fraction of the hyperparameter space [16, 17]. Metaheuristic algorithms offer dynamic and non-linear search strategies that can escape local optima and reduce computational cost [18]. The GWO is a population-based metaheuristic inspired by the social hierarchy and hunting strategies of grey wolves. Wolves are modeled in hierarchical roles (alpha, beta, delta, and omega), and the algorithm simulates encircling prey, searching for, and attacking the target [19]. Metaheuristic optimization has been paired with ML models in various domains [20], including prediction of efficiency in information technology firms [21], disease analysis [22], feature selection [23], and parameter identification of photovoltaic models [24]. Such results suggest that integrating metaheuristic algorithms with ML improves forecasting accuracy, yet this combination remains largely unexplored in macroeconomic forecasting.

Another challenge for ML-based forecasting is interpretability. Policymakers need to understand which variables drive predictions to evaluate their plausibility and design interventions. The opaque nature of complex models can reduce trust and hinder adoption. Techniques such as SHAP provide post-hoc interpretability by attributing contributions to individual features. Growth theory has long acknowledged that institutions and governance are fundamental determinants of economic performance. Yet experimental results on the connection between governance and growth are mixed. Some studies find no association between government effectiveness and subsequent growth. Others, using larger datasets and dynamic estimators, find that government effectiveness has a substantial positive effect on the growth of economy [13]. Despite this evidence, governance indicators rarely appear in machine-learning forecasts of GDP. There is therefore an unaddressed opportunity to examine whether incorporating government effectiveness into ML-based forecasting can improve accuracy and generate insights about institutional drivers of growth.

While many studies have used machine learning methods for macroeconomic forecasting, three areas remain of concern. The first area is the lack of studies that use metaheuristic-optimized XGBoost. Metaheuristic-optimized XGBoost has shown improved results in several other studies, including energy optimization [25], sales forecasting [26],

and economic-cycle classification [27]. However, the use of metaheuristic-optimized XGBoost in GDP-level macroeconomic forecasting has not been explored. Most of the previous studies used XGBoost but did so without parameter optimization [28]. None of these techniques leverages the population-based, gradient-free search characteristics of bio-inspired metaheuristics to optimize the large number of hyperparameters associated with XGBoost. While this does not mean that prior models were poorly optimized, the unique combination of Grey Wolf Optimizer (GWO) cooperative search mechanism and XGBoost's regularized boosting has not been investigated in this forecasting domain. The second area of concern is the lack of institutional quality in ML-based GDP forecasting models. Institutional quality is generally recognized as a structural determinant of long term economic performance [13, 14] and therefore should be included in any GDP forecasting model. A recent study [29] employed governance indicators alongside trade network features in an ML framework for growth prediction; however, this study did not test metaheuristic-tuned models nor provide game-theoretic explanations of feature interactions. Other GDP forecasting studies utilizing LSTMs [2, 29], Random Forests [30, 31], and Gradient Boosting [33] limited themselves to traditional macroeconomic indicators such as industrial production, unemployment, trade, etc. These studies did not explore the predictive power of governance quality. Finally, while several recent studies have utilized SHAP to explain economic forecasting models [33, 34] and metaheuristic algorithms have been coupled with XGBoost in numerous applications [19, 35, 36, 37]; as far as we know no study currently exists that integrates all three aspects: automatic hyperparameter selection through a population-based metaheuristic, inclusion of institutional quality into the set of predictive features and post hoc explanation by means of SHAP. These three missing elements limit both the accuracy of current GDP forecasts and their policy relevance. The contributions of this study is summarised as follows:

1. In this study we present an integrated GWO-XGBoost to forecast the Monthly National Level Gross Domestic Product (GDP).
2. Government effectiveness is incorporated with traditional macroeconomic variables.
3. The proposed integrated model is compared to other Machine Learning architectures.
4. SHAP Explainer is implemented used to generate Post-Hoc feature attributions from the trained model.

The rest of the research is structured as follows: In Sect. 2 we detail previous research; Sect. 3 provides the methodology of the proposed hybrid model. This includes the Original GWO, XGB-boost Algorithm, Proposed hybrid model, data used, evaluation metrics; In Sect. 4 we outline the experiments and discussion; and finally in Sect. 5 we detail the Conclusion.

2 Related work

2.1 Machine learning architectures for GDP forecasting

The application of machine learning to GDP forecasting has progressed from simple neural networks to sophisticated deep learning and ensemble architectures. Ngoc et al. [30] developed a phase-adaptive attention LSTM (PAA-LSTM) that adjusts attention weights based on the stage of the economic cycle. By aligning the learning process with the recession, recovery, expansion, and stagnation phases, their model captures cyclical dynamics that are missed by standard deep learning approaches. When tested on GDP,

the PAA-LSTM consistently achieved lesser root mean square errors and mean absolute errors than bi-directional LSTM. Oancea [38] examined the predictive capability of univariate LSTM networks for quarterly GDP and compared them with a classical SARIMA model. The LSTM captured nonlinearities and produced lower mean absolute percentage error, mean absolute error, and mean squared error than SARIMA across all forecast horizons. Mourtas et al. [39] proposed a multi-function activated weight and structure determination time-series neural network (MWASDT) for GDP prediction. The model introduces novel activation functions that adapt to different data regimes and automatically determine the number of hidden neurons. Experiments showed that MWASDT outperformed traditional neural networks and machine-learning models on GDP datasets, particularly on unseen data, where it maintained low prediction error. The automatic architecture selection reduced the need for manual hyperparameter tuning. Jallow et al. [40] evaluated generative adversarial networks for forecasting the Gambia's GDP. They trained GANs to generate realistic GDP trajectories and compared their performance against RF and XGBoost models.

The GAN achieved better predictive accuracy. Xia [33] proposed a combined predictor that fuses linear regression and XGBoost regression through entropy-based weighting to forecast regional GDP. The entropy method assigns higher weights to models with lower forecast errors, and the combined prediction is a weighted average of individual forecasts. Their experiment on Chinese provincial GDP showed that the ensemble model achieved better performance. The study shows that combining linear and tree-based learners can improve accuracy without complex tuning. Passarella et al. [31] performed a comparative evaluation of machine-learning algorithms for predicting Indonesia's GDP growth. They applied random forest, XGBoost, Support Vector Regression (SVR), Huber Regressor (HR), Decision Tree (DT), and Multi-Layer Perceptron (MLP) within a fuzzy logic framework. Among these, the RF delivered the best out-of-sample accuracy when predicting GDP. Jallow et al. [41] explored a hybrid RNN-LSTM framework with transfer learning to model GDP growth in the Gambia. The model leverages parameter transfer from pre-trained networks and optimizes hyperparameters through random search. It achieved a better accuracy and low root mean square error, outperforming standalone recurrent networks. Soybilgen and Yazgan [32] investigated the prediction of U.S. GDP by combining dynamic factor models with Bagged Trees (BT), RF, and GB. GB was augmented by incorporating dynamic, time-varying factors into the model to improve forecasting performance over traditional linear models.

However, despite significant advancements in recent years, there remain major issues inherent in the large body of research on predicting GDP: the limited exploration of metaheuristic parameter optimization of the models, while other optimization strategies demonstrated improvement.

2.2 Metaheuristic-optimized machine learning in forecasting applications

A parallel stream of literature has demonstrated the value of metaheuristic algorithms for hyperparameter optimization in forecasting contexts within macroeconomics. Chen and Wu [42] introduced a chaotic whale optimization algorithm (CWOA) to tune a Back-Propagation Neural Network (BPNN) and then combined it with an Autoregressive Integrated Moving Average (ARIMA) model to predict Chongqing's GDP. LASSO regression first selected the most relevant macroeconomic indicators, then the CWOA

searched for optimal neural network weights, and the ARIMA modeled residual errors. Their hybrid CWOA-BPNN-ARIMA system reduced mean absolute error and root mean square error compared with an RF baseline. Sun et al. [27] employed a multi-objective Lichtenberg algorithm (MOLA) with an XGBoost model for forecasting China's economic cycle, and the results demonstrated better accuracy. However, many areas remain unexplored. Specifically, the MOLA-XGBoost study used cyclic categorization rather than predicting continuous GDP values, and it did not include governance indicators. While Chen and Wu's framework is a metaheuristic optimizer, the authors did not employ an ensemble model as their base learning algorithm and did not conduct an interpretability analysis.

2.3 Summary of gaps and positioning of current research

The preceding survey of the existing literature identified three related gaps intended to be addressed in this research. First, although some researchers have shown that using a metaheuristic to optimize the hyperparameters of an XGBoost machine learning model can yield better predictive performance than manually tuning them, no prior study has applied GWO-XGBoost to monthly national-level GDP forecasting for the US. Second, none of the previously published machine-learning-based GDP forecasts included governance quality as a predictor, leaving a significant gap. Finally, no known methodology combines a metaheuristic optimization algorithm with SHAP-based feature importance explanations, while also including governance-quality variables, into a single GDP forecasting method. Thus, the proposed GWO-XGBoost with SHAP framework bridges these three gaps. It combines automated hyperparameter search, institution-augmented feature modeling, and game-theoretic post-hoc attribution explanations into a single coherent, policy-oriented forecasting tool. The placement of this proposed framework is summarized in the comparison table presented in Table 1.

3 Research methodology

3.1 Grey Wolf Optimizer (GWO)

The GWO is a population-based metaheuristic algorithm inspired by the social hierarchy and cooperative hunting strategies observed in grey wolf packs [47]. Central to the algorithm's design is a dominance structure comprising four roles: alpha (α), beta (β), delta (δ), and omega (ω) [48]. The α individual represents the pack leader and is considered the optimal solution at any given iteration; the β and δ wolves support decision-making and act as secondary and tertiary guides, respectively; the remaining members-designated as ω -follow these leaders during the search process. The GWO algorithm models three primary phases of predation: (1) tracking and approaching prey, (2) encircling and harassing the target, and (3) attacking once the prey is immobilized. These stages collectively emulate the algorithm's transition from global exploration to local exploitation. The mathematical formulation of GWO begins with the random initialization of potential solutions (wolf positions) inside the bounded problem space as defined in Eq. (1).

$$\mathrm{X} = (\mathrm{UB} - \mathrm{LB}) \cdot \phi + \mathrm{LB} \tag{1}$$

where UB and LB denote the upper and lower bounds of the decision variables, ϕ is a uniformly distributed random vector in $[0,1]$, and X represents the position vector of

Table 1 Comparative taxonomy of related studies and the present contribution

#	Reference	Year	Application	Base model	Parameter optimization	Interpretability	Frequency	Key limitations
1	Chen and Wu [42]	2023	GDP forecasting	BP-NN with ARIMA	Chaotic WOA	None	Monthly	No SHAP analysis
2	Xia [33]	2023	GDP forecasting	Linear Regression with XGBoost fusion	Entropy weighting	None	Annual	No SHAP analysis; No Government Effectiveness variable
3	Mourtas et al. [39]	2023	GDP forecasting	MWASDT-NN	Auto architecture	None	Quarterly	No SHAP analysis; No Government Effectiveness variable
4	Silva et al. [29]	2024	GDP forecasting	RF, XGBoost, LightGBM	None	SHAP	Monthly	No metaheuristic optimization
5	Kamruzzaman et al. [43]	2024	GDP Prediction	LR, RF, XGBoost, LightGBM	None	SHAP	Quarterly	No metaheuristic optimization; No Government Effectiveness variable
6	Jallow et al. [40]	2025	GDP forecasting	GAN	Grid Search	None	Annual	No SHAP analysis; No Government Effectiveness variable
7	Oancea [38]	2025	GDP forecasting	Univariate LSTM	None	None	Quarterly	Univariate; No SHAP analysis; No metaheuristic optimization;
8	Passarella et al. [31]	2025	GDP forecasting	RF	None	SHAP	Annual	No metaheuristic optimization; No Government Effectiveness variable
9	Ngoc et al. [30]	2025	GDP forecasting	Phase-adaptive LSTM	Phase Wise Tunning	None	Annual	No SHAP analysis; No Government Effectiveness variable
10	Sun et al. [27]	2025	Economic cycle forecasting	MOLA-XGBoost	MOLA	SHAP	Quarterly	No Government Effectiveness variable
11	Akepanidtaworn and Akepanidtaworn [44]	2025	GDP nowcasting	Machine Learning Algorithms	None	None	Quarterly	No metaheuristic; No SHAP; No Government Effectiveness variable
12	Raghavendran et al. [45]	2026	GDP prediction	Lightweight-NN	grid search	SHAP	-	No metaheuristic; No Government Effectiveness variable;
13	Hisam et al. [46]	2026	GDP forecasting	ARIMA, Exponential Smoothing, XGBoost	None	None	Annual	No metaheuristic; No SHAP; No Government Effectiveness variable
14	This study	2026	GDP forecasting	XGBoost	GWO	SHAP	Monthly	Addresses all three gaps

an individual in the d-dimensional search space. After initialization, the fitness value of each search agent is computed. The three individuals with the highest fitness values are designated as α, β, and δ, respectively. Subsequently, the positions of all other wolves are updated relative to these leaders using the following equations in Eqs. (2–4)

$$\begin{cases} \mathrm{D}_\alpha = |\mathrm{C}_1 \cdot \mathrm{X}_\alpha - \mathrm{X}| \\ \mathrm{D}_\beta = |\mathrm{C}_2 \cdot \mathrm{X}_\beta - \mathrm{X}| \\ \mathrm{D}_\delta = |\mathrm{C}_3 \cdot \mathrm{X}_\delta - \mathrm{X}| \end{cases} \tag{2}$$

$$\begin{cases} \mathrm{X}_1 = \mathrm{X}_\alpha - \mathrm{A}_1 \cdot \mathrm{D}_\alpha \\ \mathrm{X}_2 = \mathrm{X}_\beta - \mathrm{A}_2 \cdot \mathrm{D}_\beta \\ \mathrm{X}_3 = \mathrm{X}_\delta - \mathrm{A}_3 \cdot \mathrm{D}_\delta \end{cases} \tag{3}$$

$$\mathrm{X}(t+1) = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \mathrm{X}_3}{3} \tag{4}$$

In these expressions, X_α, X_β, and X_δ denote the current positions of the three leading wolves. $\mathrm{A}_1, \mathrm{A}_2$, and A_3 and $\mathrm{C}_1, \mathrm{C}_2$, and C_3 are coefficient vectors defined as expressed in Eqs. (5) and (6)

$$\mathrm{A} = 2a \cdot \mathrm{r}_1 - a \tag{5}$$

$$\mathrm{C} = 2\mathrm{r}_2 \tag{6}$$

where r_1 and r_2 are random vectors in $[0,1]$, and a is a linearly decreasing parameter that controls the transition from exploration to exploitation, its mathematical expressed in Eq. (7)

$$a = 2 - \frac{2t}{M} \tag{7}$$

Here, t represents the current iteration, and M is the maximum number of iterations. At the beginning of the search ($t = 0$), $a = 2$, promoting broad exploration. As the algorithm progresses, a declines to zero, intensifying local search around promising regions. The coefficient vectors A and C modulate the divergence from (when $|\mathrm{A}| > 1$) or convergence toward (when $|\mathrm{A}| < 1$) the leaders, thereby enabling adaptive balance between exploration and exploitation. After position updates, the fitness values are reevaluated, and the top three solutions are reassigned as the new α, β, and δ. This iterative cycle continues until a termination criterion-typically the maximum number of iterations, is satisfied. Ultimately, the algorithm returns the position of the α wolf as the best approximation to the global optimum.

3.2 Extreme gradient boosting (XGBoost)

The XGBoost, proposed by Chen and Guestrin [48, 49], is an advanced implementation of the gradient boosting decision tree framework that leverages second-order optimization techniques to enhance model performance and generalization. Grounded in the principles of classification and regression trees, XGBoost is particularly adept at handling complex, non-linear relationships in both regression and classification tasks. Its efficacy stems from a carefully balanced objective function that jointly minimizes prediction error and controls model complexity through explicit regularization. At each iteration of the training process, XGBoost incrementally constructs a decision tree aimed at

reducing the residual errors of the ensemble model built up to that point. Unlike conventional boosting methods that rely solely on first-order gradients, XGBoost employs a second-order Taylor expansion of the loss function to more accurately approximate its curvature, thereby enabling more stable and efficient parameter updates. To prevent overfitting, a regularization term is incorporated into the objective function, which penalizes both the number of leaf nodes and the magnitude of their associated scores. The objective function minimized at each boosting round is expressed as in Eq. (8)

$$L(\phi) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \Omega(f) \tag{8}$$

where y_i denotes the true target value for the i-th observation, $\hat{y}_i$ represents the corresponding model prediction, and $\ell(\cdot)$ is a differentiable loss function-commonly the squared error in regression contexts, is defined in Eq. (9)

$$\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{9}$$

The regularization component $\Omega(f)$ is defined as given in Eq. (10)

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{10}$$

T signifies the total number of terminal (leaf) nodes in the tree, w_j denotes the score (or weight) assigned to the j-th leaf, and γ and λ are hyperparameters that govern the complexity penalty: γ controls the cost of introducing additional leaves, while λ penalizes large leaf weights to encourage smoother predictions.

3.3 Proposed GWO-XGBoost

The analytical methodology runs sequentially as follows: the first step involves collecting all data relevant to this study, cleaning and preprocessing each dataset, and applying min-max normalization to remove the effects of different units on the results (Sect. 3.5). In the second step, an evolutionary search over a six-dimensional hyper-parameter space is conducted using the Grey Wolf Optimizer to find the best possible configuration for the XGBoost model. For this purpose, the Mean Squared Error (MSE) on the training set is used as the objective function. The third step involves training the best possible configuration for the XGBoost model identified in step 2 using the training portion of the data and comparing its performance with that of several machine learning models using five evaluation criteria (Sect. 3.4). The fourth step uses SHAP values to analyze how each input variable contributes to the GDP prediction. By doing so, we can provide policymakers with information on the factors that contribute most significantly to a country's GDP levels (Sect. 4.2).

In this research, we selected the GWO for several key reasons. First, GWO is relatively simple and has only a few hyperparameters, both of which are important characteristics when optimizing machine learning algorithms. This is because optimizing too many hyperparameters can create instability. Second, GWO naturally shifts from the initial global search phase to an exploitation phase by adjusting its control parameter as it progresses. This allows the search to be initially very broad and then narrow down as it finds promising areas. Finally, previous studies have shown that GWO performs well for tuning the hyperparameters of ML models. Therefore, based on both theoretical

and empirical results, we chose GWO as our optimizer of choice for finding optimal combinations of the large number of nonlinear and interactive hyperparameters used in XGBoost [46, 50]. The decision to choose XGBoost as the primary predictive model was driven by a number of factors. First, XGBoost has been shown to be highly effective at modeling both linear and nonlinear relationships among predictor variables. Secondly, the regularization built into the XGBoost framework provides an additional tool to mitigate overfitting, as XGBoost uses a boosting approach to add new trees to the model. Finally, the tree-based nature of XGBoost and the ability to train models in parallel make the computational costs associated with training many different versions of the model during hyperparameter search much lower than they would be with other types of models. Prior studies have also demonstrated that tree-based ensemble methods tend to outperform other machine learning approaches in forecasting applications [51, 52, 53].

In order to overcome the fundamental problem of configuring hyperparameters for XGBoost, this research introduces an innovative hybrid optimization approach using GWO's metaheuristic search capabilities to optimize critical hyper-parameters of XGBoost. The combination has been developed to resolve some disadvantages associated with hand-tuning or grid-search based on hyper-parameters, such as being either time-consuming or failing to achieve global optimum. The proposed methodology capitalizes on GWO's biologically inspired social hierarchy and cooperative hunting dynamics to navigate the complex, non-convex parameter space of XGBoost with enhanced efficiency and robustness. The core innovation lies in treating the hyperparameter tuning process as a continuous optimization problem. Each individual within the GWO population represents a unique candidate set of XGBoost hyperparameters. The parameters of XGBoost were tuned based on predefined search ranges specified before model training. For the proposed framework, GWO was used to iteratively search within these bounds and identify the hyperparameter combination that produced the best predictive performance according to the selected objective function. The operational workflow of GWO-XGBoost framework is illustrated in Fig. 1. The algorithm commences by initializing a population of N grey wolves, where each wolf's position vector X is a 6-dimensional point in the defined hyperparameter space, generated according to Eq. (1). In each iteration, the fitness of every individual in the population is evaluated. Crucially, the fitness function is defined as the predictive performance of an XGBOOST model trained using the specific hyperparameter set encoded by that individual. The primary performance metric used for fitness assessment is the Mean Squared Error (MSE), A lower MSE value signifies a better fitness score, indicating a more optimal hyperparameter configuration. Following the fitness evaluation, the three individuals with the best fitness scores are selected as the alpha (α), beta (β), and delta (δ) wolves. The remaining wolves (ω) then update their positions in the hyperparameter space by moving towards these elite leaders, guided by the mathematical models outlined in Eqs. (2)-(4).

This collaborative movement mechanism ensures that the entire population converges toward promising regions of the search space while preserving a degree of diversity through the stochastic nature of the coefficient vectors A and C. The convergence behavior is dynamically regulated by the linearly decreasing parameter a, as defined in Eq. (7). In the initial stages of the optimization, a is high (starting at 2), promoting broad exploration across the hyperparameter landscape. As iterations progress, a diminishes towards zero, shifting the algorithm's focus to intensive exploitation around the most

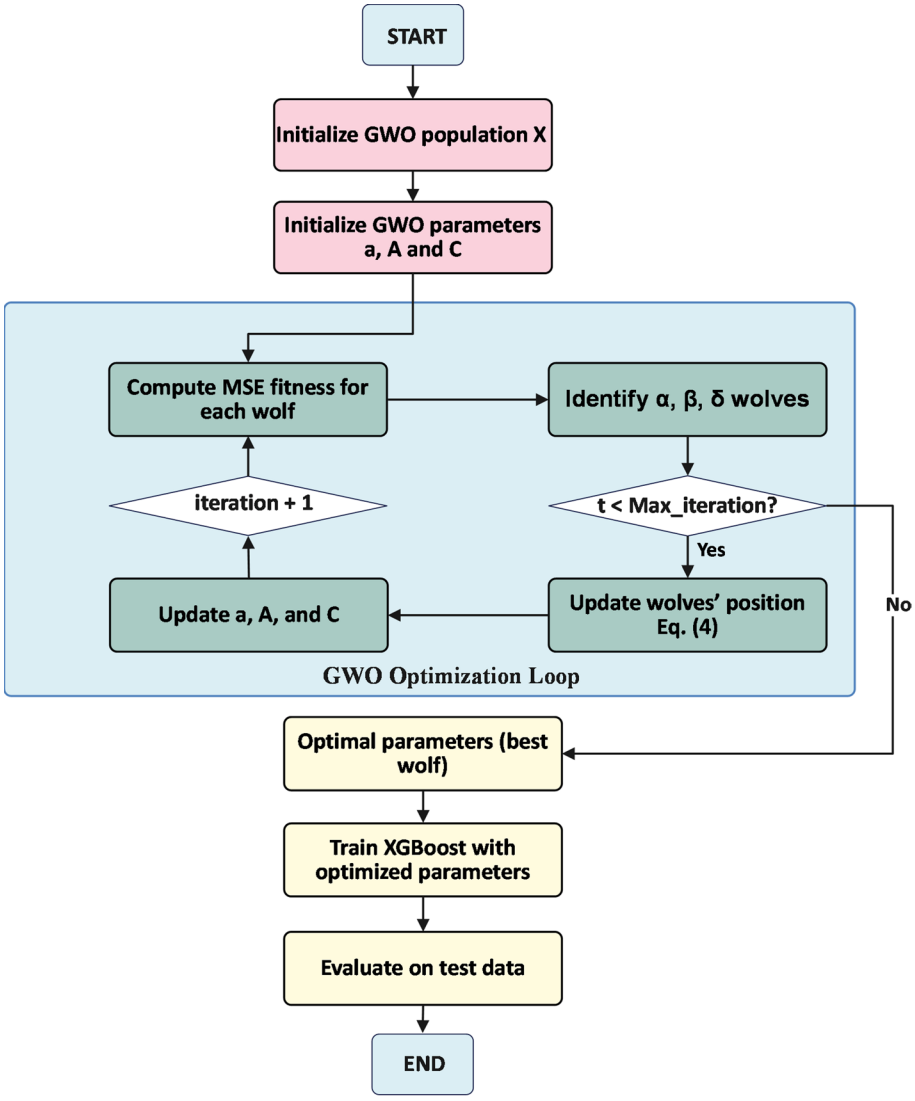


Fig. 1 Proposed GWO-XGBOOST model

promising configurations identified by the α, β, and δ wolves. This adaptive balance between exploration and exploitation is fundamental to the algorithm's ability to escape local optima and locate the global optimum. The iterative process continues until a predefined termination criterion is met-typically, the maximum number of iterations M is reached. Upon convergence, the hyperparameter set corresponding to the final α-wolf position is selected as the optimal configuration for the XGBoost model. This optimized model is then retrained on the full training dataset and deployed for prediction on the hold-out test set to evaluate its final performance. This proposed GWO-driven hyperparameter optimization strategy offers a systematic, automated, and highly effective approach to configuring XGBoost. By harnessing the collective intelligence and adaptive search mechanisms of GWO, the framework significantly improves the accuracy and generalization capacity of the underlying XGBoost model, thereby providing a powerful tool for tackling complex regression and classification tasks in data science applications. Table 2 presents the search bounds for the six XGBoost hyperparameters optimised by the Grey Wolf Optimizer in this study. The bounds for each hyperparameter were

Table 2 GWO search space for XGBoost hyperparameter optimization

Hyperparameter	Symbol	Lower Bound (LB)	Upper Bound (UB)	Description
Learning rate	lr	0.001	0.05	Step size shrinkage to prevent overfitting
Maximum tree depth	max_depth	3	15	Maximum depth of each decision tree
Number of boosting rounds	n_estimators	10	200	Total number of sequential trees in the ensemble
Subsample ratio	subsample	0.5	1.0	Fraction of training instances sampled per tree
Column subsample ratio	colsample_bytree	0.5	1.0	Fraction of features sampled per tree
Regularisation penalty	gamma	0.0	5.0	Minimum loss reduction required for a further partition

chosen based on two complementary criteria. First, the ranges align with values commonly reported in prior metaheuristic-optimised XGBoost studies in forecasting applications, ensuring that the search space is consistent with the empirical norms of the field. Second, the bounds reflect a deliberate trade-off between computational overhead and search efficiency: each GWO iteration evaluates the fitness of every wolf in the population by training a complete XGBoost model on the training partition, so the cumulative cost of the metaheuristic loop scales with both the population size and the volume of the search space. Narrower bounds reduce per-iteration cost but risk excluding promising regions of the hyperparameter space, while excessively wide bounds inflate the search cost without proportional gains in predictive accuracy. The bounds reported in Table 2 balance these two considerations and ensure that the optimisation remains tractable on the computational resources available for this study.

3.4 Evaluation metrics

The predictive capability of the GWO optimized XGBoost model is assessed using a set of multiple error and accuracy measures in order to evaluate each component of the model's performance (bias, variance, relative deviation) as well as the models' overall performance. The evaluation indicators used include the coefficient of determination (R^2), root mean squared error (RMSE), mean squared error (MSE), max error (ME), and mean bias error (MBE). Their mathematical formulations are as follows in Eqs. (11–15)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{12}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

$$ME = \max_i (|y_i - \hat{y}_i|)^2 \tag{14}$$

$$MBE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \tag{15}$$

where y_i and $\hat{y}_i$ represent the actual and predicted values, $\bar{y}$ denotes the mean of the observed values respectively, and n denotes the total number of observations. Together, these metrics offer a multidimensional perspective on model accuracy, robustness, and reliability.

3.5 Data

This study utilizes a comprehensive macroeconomic dataset for the United States spanning the period from January 1996 to December 2020 at a monthly frequency. Because these variables were originally reported at annual frequency, they were converted to monthly frequency in EViews 14 to expand the sample size. This preprocessing step was important for machine learning implementation, since a larger number of observations improves training stability, supports pattern extraction, and reduces the limitations associated with sparse annual data. The selected indicators collectively capture key dimensions of economic activity, labor market conditions, trade dynamics, consumer behavior, institutional quality, and overall economic output. Specifically, the dataset comprises the following six variables: (1) the Industrial Production Index (INDPRO), sourced from the Federal Reserve Economic Data (FRED) repository [54]; (2) the Unemployment Rate (UNEMP) obtained from the World Bank Database [55]; (3) the Trade Balance (TBAL) for Goods and Services [54]; (4) Advance Retail Sales for Retail Trade (RSALES) [54]; (5) the Government Effectiveness index (GOVEFF), drawn from the World Bank Database [55]; and (6) Gross Domestic Product (GDP) [55]. The target variable is GDP, while other variables are the input variables. Crucially, to mitigate the adverse effects of scale disparity and to facilitate machine learning void of bias, all features including the target variable (GDP) were normalized using the min-max scaling transformation as defined in Eq. (16).

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

where x denotes the original value of a variable, and $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values observed for that variable. This normalization maps all values to the closed interval $[0,1]$, while maintaining the relative relationships among observations while eliminating unit-dependent biases that could disproportionately influence model learning. The correlation matrix and the data distribution of each variable are given in Figs. 2 and 3, respectively. Table 3 shows variable's definition, source, and theoretical justification.

4 Experiment and discussion

4.1 Performance comparison

To ensure a rigorous and fair comparative assessment, this study benchmarks the proposed GWO-XGBoost framework against established machine learning models: the standard XGBoost, an MLP, SVM, RF, Light Gradient Boosting Machine, and DT. Their parameters are given in Table 4. The GWO was employed with a population size of 30 search agents and a maximum of 50 iterations. All experiments adopted a random 80:20 training-to-testing data split ratio. The selection of baseline models is guided by several criteria designed to ensure both methodological rigor and coverage of the contemporary forecasting landscape. First, architectural diversity: the final pool spans, boosting

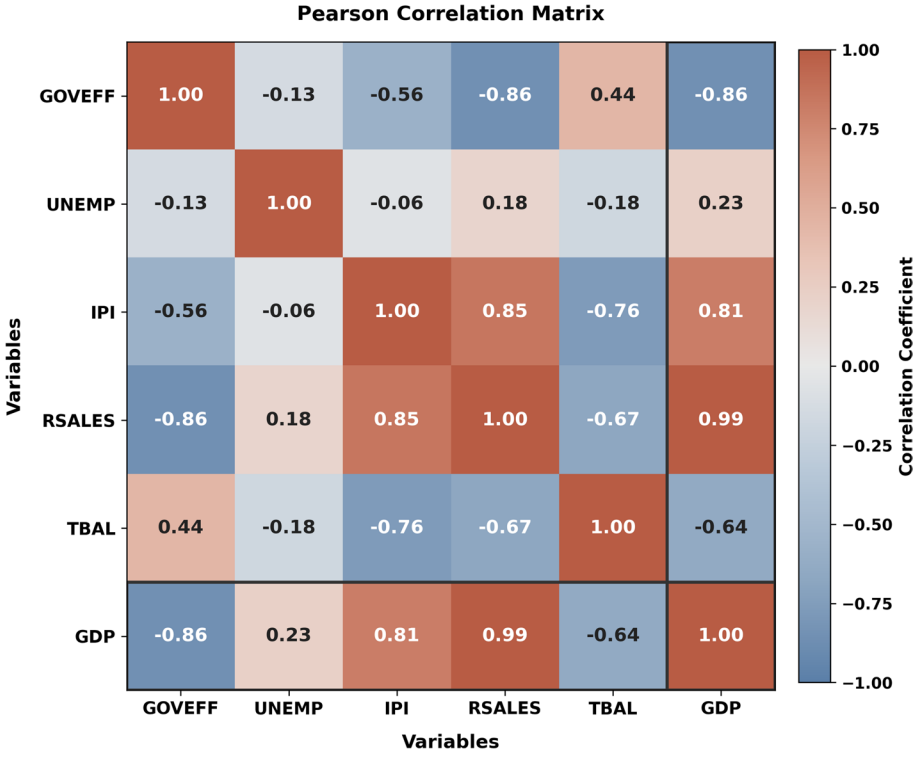


Fig. 2 Correlation matrix of variables

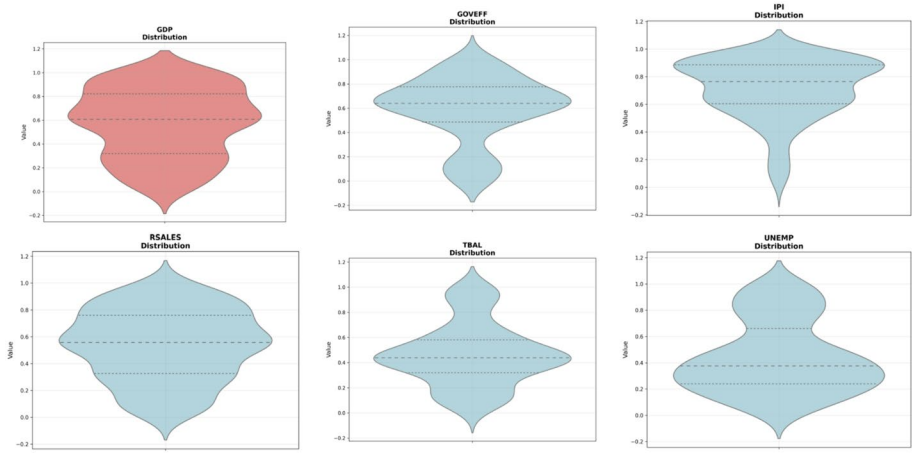


Fig. 3 Distribution of variables

ensembles, neural networks, and kernel methods, ensuring that the proposed GWO-XGBoost framework is tested against a representative range of learning paradigms.

The empirical evaluation of the proposed GWO-XGBoost framework was conducted to rigorously assess its predictive efficacy against established machine learning models. The performance comparison was performed on both the training and test datasets, with results quantified across five distinct metrics. As presented in Table 5, the GWO-XGBoost model demonstrated improved performance across all evaluated metrics on the training set. It achieved an R^2 score of 0.99127, significantly outperforming the standard XGBoost, MLP and SVM models. This indicates that the optimized model

Table 3 Variable definition and theoretical justification

Variable	Definition	Source	Theoretical justification
GDP	Gross Domestic Product (current USD)	World Bank [55]	Target variable; aggregate measure of national economic output
RSALES	Advance Retail Sales for Retail Trade	FRED [54]	Proxy for consumer demand
INDPRO	Industrial Production Index	FRED [54]	Proxy for production-side economic activity; captures manufacturing, mining, and utilities output aligned with the supply-side of national accounting
UNEMP	Unemployment Rate	World Bank [55]	Labour market indicator
TBAL	Trade Balance for Goods and Services	FRED [54]	Proxy for net exports; the external sector component of the GDP expenditure.
GOVEFF	Government Effectiveness Index	World Bank [55]	Proxy for institutional quality; institutional economics theory identifies governance as a structural determinant of long-term growth through incentive structures, property rights enforcement, and policy credibility

Table 4 Compared machine learning model parameters

Model	Hyperparameter configuration
XGBoost	learning_rate = 0.01; max_depth = 15; n_estimators = 200;
MLP	Hidden layers = (100, 50); activation = ReLU; optimiser = Adam; learning_rate = 0.001; epochs = 500
SVM	kernel = RBF; C = 100.0; γ = 'scale'; epsilon = 0.1
DT	max_depth = 15, min_samples_split = 3

Table 5 Performance comparison of models on the training set

Model	R^2	RMSE	MSE	ME	MBE
XGBoost	0.97698	0.15172	0.02302	0.45120	0.00053
MLP	0.97620	0.15429	0.02381	0.41699	-0.01572
SVR	0.96449	0.18843	0.03551	0.30032	0.06407
GWO-XGBoost	0.99127	0.09345	0.00873	0.31093	0.00070

explains over 99% of the variance in the training data. The substantial improvement in R^2 is corroborated by a marked reduction in error metrics: the RMSE decreased from 0.15172 (XGBoost) to 0.09345, and the MSE plummeted from 0.02302 to 0.00873. These reductions signify not only a lower average magnitude of prediction errors but also a drastic decrease in the spread of those errors, suggesting enhanced stability and precision in the model's forecasts. Furthermore, the ME values for GWO-XGBoost were 0.31093, respectively, indicating that while the maximum error remains moderate. The MBE value of 0.00070 reveals negligible systematic bias, confirming that the model's predictions are neither consistently overestimating nor underestimating the true values during training. In contrast, the MLP exhibited a notable negative bias, while the SVM displayed a positive bias, highlighting the robustness of the GWO-tuned model in achieving unbiased predictions.

The true test of any predictive model lies in its ability to generalize to unseen data. As shown in Table 6, the GWO-XGBoost model maintained its dominant performance on the test set, achieving an R^2 of 0.99114, which is nearly identical to its training performance. This minimal degradation underscores the model's exceptional generalization capacity and suggests it is not suffering from overfitting, a common pitfall in complex models. The RMSE and MSE on the test set were 0.09468 and 0.00897, respectively, again representing a significant improvement over the baseline XGBoost and

Table 6 Performance comparison of models on the test set

Model	R^2	RMSE	MSE	ME	MBE
XGBoost	0.97635	0.15471	0.02394	0.36070	-0.00487
MLP	0.96346	0.19232	0.03699	0.43357	-0.03869
SVM	0.95756	0.20726	0.04296	0.35642	0.03957
GWO-XGBoost	0.99114	0.09468	0.00897	0.24592	-0.00344

the other competitors. The model's low MBE of -0.00344 on the test set further confirms its calibration, demonstrating that the slight negative bias observed during training does not amplify in the out-of-sample data. The ME of 0.24592 indicates that the model's predictions are not only accurate but close to the observed value in the worst case. The MLP and SVM models, while still performing reasonably well, lagged considerably behind, with RMSEs exceeding 0.19 and R^2 scores below 0.96, underscoring their inferior capacity to capture the underlying patterns in the data. The experimental results unequivocally validate the hypothesis that hyperparameter optimization via the GWO significantly enhances the predictive power of the XGBoost algorithm. The GWO-XGBoost model consistently outperformed all comparative methods across every metric and on both training and test partitions. This superiority can be attributed to the GWO's ability to navigate the complex, multi-dimensional hyperparameter space effectively, identifying configurations that balance model complexity and predictive accuracy, a task that manual tuning or grid search often fails to accomplish efficiently. The R^2 scores on both sets suggest that the model has successfully captured the fundamental relationships within the macroeconomic indicators used as predictors. The consistent reduction in RMSE and MSE, coupled with the minimal MBE, indicates that the GWO-optimized model produces forecasts that are not only precise but also unbiased, making it particularly suitable for applications where directional accuracy is critical. The negligible difference between training and test performance further attests to the model's robustness and reliability for real-world deployment. In conclusion, the integration of GWO for hyperparameter tuning represents a significant methodological advancement for XGBoost-based forecasting systems. It transforms a powerful yet potentially suboptimal model into one capable of achieving state-of-the-art performance, thereby providing researchers and practitioners with a highly effective tool for predictive analytics in economic and financial domains.

To provide a qualitative and intuitive assessment of model fidelity, Fig. 4 presents scatter plots comparing predicted values with actual observations for the evaluated models XGBOOST, MLP, SVM, and the proposed GWO-XGBoost on both the training and test datasets. These plots serve as a powerful complement to the quantitative metrics presented in Tables 5 and 6, offering direct insight into the distribution of prediction errors, the presence of systematic bias, and the overall alignment between forecasts and ground truth. The scatter plot for the baseline XGBoost model reveals a generally strong positive correlation between predicted and actual values, consistent with its high R^2 score of 0.976 on the training set. The data points cluster closely around the diagonal line of perfect agreement, particularly in the mid-range of the target variable. However, a slight deviation from the ideal line is observable at the lower and upper extremes, suggesting minor underestimation or overestimation in these regions. On the test set, the dispersion appears marginally wider than in the training set, indicating some degree of

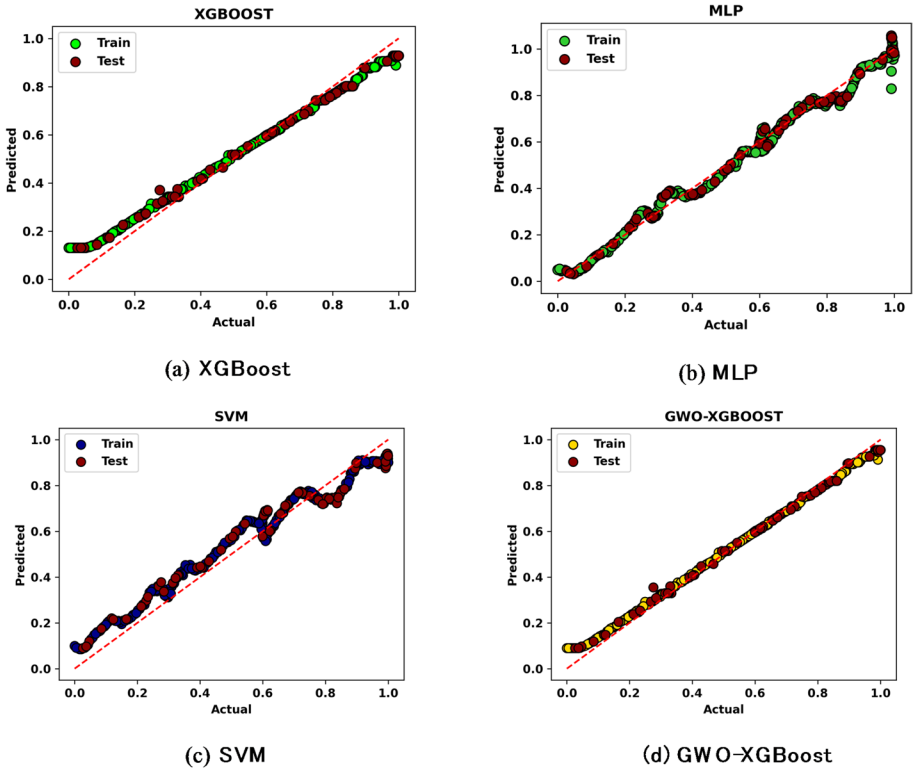


Fig. 4 Scatter plots of predicted vs. actual GDP for all models on training and test sets

performance degradation during out-of-sample prediction, a phenomenon that is quantitatively reflected in its slightly lower test-set R^2 of 0.97635.

The MLP's scatter plot demonstrates a robust fit on the training data, with predictions tightly near the diagonal. Nevertheless, a more pronounced non-linear distortion becomes evident, especially at higher values, where the model appears to exhibit a tendency toward underprediction. This aligns with its negative MBE on the training and test sets, confirming a systematic downward bias. The test set points exhibit greater scatter than the training set, corroborating the model's reduced generalization capability relative to the optimized variants. For the SVM, the scatter plot exhibits a less uniform distribution of points, with visible clustering and gaps along the diagonal. While the model captures the general trend, it struggles to accurately predict values in the middle range, where points deviate most significantly from the ideal line. This is consistent with its comparatively lower R^2 scores and higher RMSE values. The positive MBE is also apparent, as the cloud of points appears slightly elevated above the diagonal, indicating a consistent overestimation bias across both datasets. In stark contrast, the scatter plot for the proposed GWO-XGBoost model displays an exceptional degree of alignment between predicted and actual values. Both training and test data points form a tight linear band along the diagonal, with minimal deviation across the entire range of the target variable. The tight clustering of points, even at the extremes, underscores the model's superior capacity to capture complex, non-linear relationships without introducing significant bias. The negligible MBE values are visually confirmed by the symmetric distribution of points around the diagonal, indicating a highly calibrated and unbiased predictive system. The near-identical positioning of training and test points further validates the

model's outstanding generalization performance, reinforcing the quantitative findings of its high R^2 scores and minimal error metrics.

These analyses confirm that while all models exhibit a fundamental ability to learn the underlying patterns in the data, the proposed GWO-XGBoost framework achieves a level of precision and fidelity unmatched by its counterparts. The plots demonstrating minimal scatter, the absence of systematic bias, and consistent performance on both the training and testing datasets provide compelling support for the efficacy of the GWO driven hyperparameter optimization strategy. This plot complements the numerical results, affirming that integrating metaheuristic search with gradient boosting not only enhances statistical accuracy but also produces a more reliable, robust predictive model suitable for practical deployment in macroeconomic forecasting applications.

4.2 Explainability analysis

4.2.1 Overall feature analysis

To move beyond purely predictive performance and elucidate the decision-making logic of the GWO-XGBoost model, a SHAP-based feature attribution analysis was conducted. SHAP provides a theoretically grounded, game-theoretic framework for attributing the model's output to its input features, ensuring local accuracy, and consistency [55, 56, 57]. Unlike permutation-based importance metrics, SHAP values quantify the marginal contribution of each feature across all possible coalitions of the remaining features, so that the sum of individual feature attributions equals the difference between a given prediction and the global average prediction.

SHAP values describe how the trained GWO-XGBoost model uses its input features to generate predicted GDP values; they do not measure the causal economic effect of any feature on actual GDP. This distinction is structural rather than cosmetic: SHAP attributes credit for the model's output, and the model's output is a function of the joint feature distribution observed in the training data. The SHAP value for feature i is computed as given in Eq. (16).

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{16}$$

where N is the full feature set, S is any subset that excludes feature i, and $f(S)$ is the model's expected output conditional on the features in S. The formulation gives a weighted-average marginal contribution of each feature across all coalitions and supports both global feature ranking and local per-observation attribution [56].

Figure 5 presents the SHAP summary plot for the five macroeconomic predictors. The horizontal axis is the SHAP value (the magnitude and direction of each feature's contribution to a single prediction), the vertical axis ranks features by mean absolute SHAP value (a global importance proxy), and the color gradient maps the underlying feature value from low (blue) to high (red). The analysis reveals a clear hierarchy of model-internal feature importance. RSALES is the dominant feature in the model's predictive structure, with the largest mean absolute SHAP value and the widest distribution. High retail sales values (red) are associated with positive SHAP contributions, while low values (blue) are associated with negative contributions. Within the model, the consumer-demand channel carries the most predictive weight. GOVEFF ranks second in the SHAP hierarchy. GOVEFF in Fig. 5 shows that high GOVEFF values (red dots) cluster on the

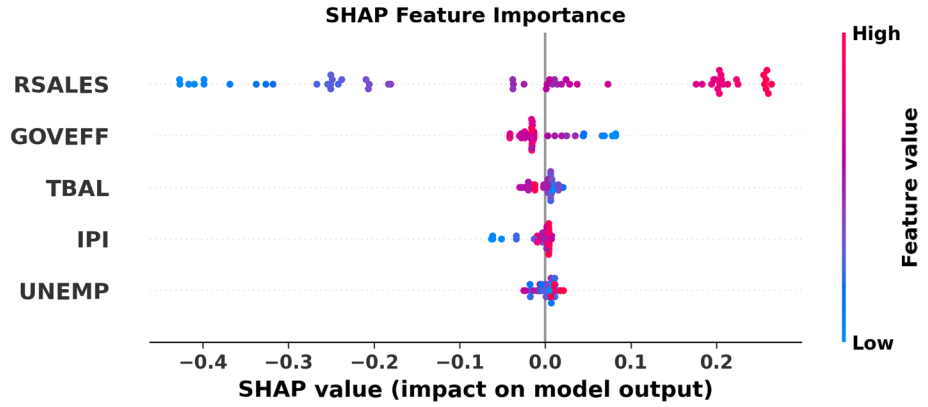


Fig. 5 SHAP analysis plot

negative side of the SHAP axis, while low GOVEFF values (blue dots) are sparsely distributed on the positive side. Within the data sample, observations with high GOVEFF values correspond to negative SHAP values, and observations with low GOVEFF correspond to positive. The model learns this association and reproduces it in the SHAP attributions: red dots (high GOVEFF, early years) carry negative contributions, and blue dots (low GOVEFF, recent years) carry positive contributions. TBAL ranks third in the SHAP hierarchy, with the majority of the values in blue associated with positive contributions and few red values with negative contributions. The model assigns meaningful but secondary weight to the net-exports channel, consistent with the broader pattern in which domestic demand indicators dominate the learned structure of the model.

IPI ranks fourth, with SHAP values spanning a relatively narrow range. Higher IPI values are weakly associated with positive SHAP contributions and lower values with slightly negative contributions. The relatively modest magnitude of IPI's SHAP contribution does not imply that industrial production is unimportant for the U.S. economy in any structural sense; it indicates that, within this specific feature set, much of the predictive signal carried by IPI is already captured by RSALES, so the model has limited additional information to extract from IPI given the other inputs. UNEMP ranks fifth, with SHAP values consistently small in magnitude. Higher unemployment values and lower values cluster around the zero axis, but the absolute magnitude of the UNEMP SHAP contribution is the smallest among the five features. Three model-and-data factors plausibly account for the pattern: collinearity between UNEMP and RSALES through the consumption–employment channel, the predominance of stable employment regimes in the data sample (which limits the variance UNEMP contributes to the model), and the absence of lagged specifications that would allow the model to capture the lead-lag dynamics of Okun's law relationships.

4.2.2 Feature dependency relationship analysis

The SHAP dependency plots in Fig. 6 show how each feature's SHAP value varies with its own value. Providing a feature-level view of the model's learned decision logic. As in Sect. 4.2.1, the patterns described below characterise the model's predictive use of each feature within the joint feature distribution. RSALES monotonic, approximately linear positive relationship. SHAP values rise steadily from approximately − 0.4 at the lowest values to + 0.25 at the highest, with very little scatter. This is the cleanest dependency

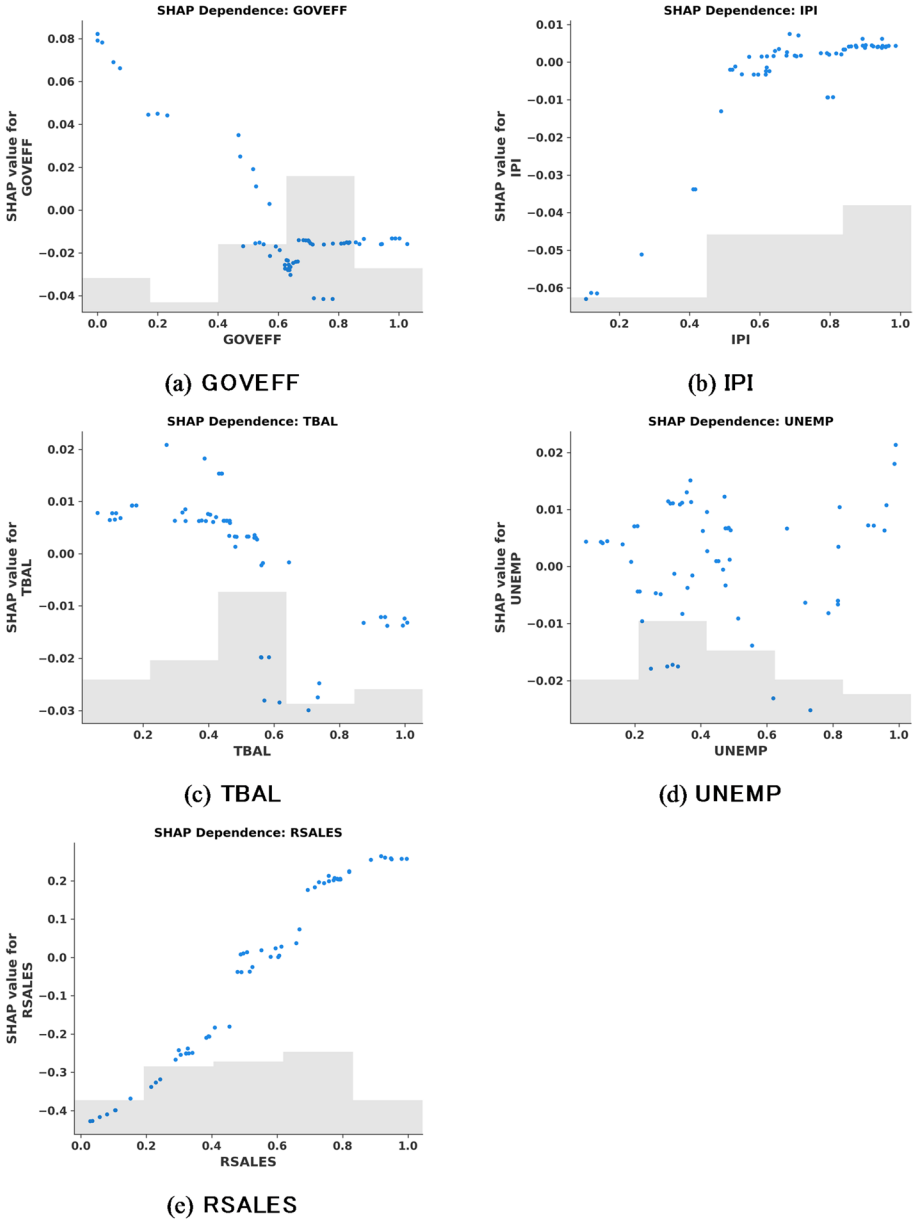


Fig. 6 Dependency plot of variables

pattern in the feature set and is consistent with the position of RSALES at the top of the global SHAP ranking. GOVEFF shows a clear non-monotonic pattern: SHAP contributions are strongly positive (up to +0.08) at low values, decline rapidly through the moderate range (0.4–0.7), and remain slightly negative beyond 0.7. As discussed in Sect. 4.2.1, this shape reflects the temporal structure of the data sample in which low GOVEFF values are concentrated in recent (high-GDP) years and high GOVEFF values in earlier (lower-GDP) years.

IPI exhibits a broadly pro-cyclical shape: SHAP values are strongly negative (≈ – 0.06) at the lowest IPI values and rise rapidly to a narrow band slightly above zero for IPI values above approximately 0.5. The magnitude of the IPI contribution is much smaller than that of RSALES. This attenuation is most plausibly explained by collinearity with RSALES, which allows the model to substitute predictive signal between two correlated

demand-side indicators. Also, the strongly negative SHAP values at low IPI are driven by a small number of points in the sparse lower tail of the distribution; tree-based ensembles produce locally unstable attributions in sparse regions. A more complete investigation of the IPI–GDP relationship would require regime-conditional or sequence-aware modeling, which will be identified as a direction for future work in conclusion section. TBAL shows a mildly negative relationship, with less scatter, and SHAP values are more positive at low TBAL. The wider scatter indicates that the model's use of TBAL depends on the values of other features in the joint distribution rather than on TBAL alone. UNEMP dependency plot shows a noisy distribution with a negative tendency, consistent with the description in Sect. 4.2.1. SHAP values fluctuate within a narrow band across the full range of the feature. The small magnitude reflects three features of the modelling setup: collinearity with the demand-side indicators (which allows the model to substitute signals between correlated features), the predominance of stable employment regimes in the dataset (which limits the variance UNEMP can contribute), and the absence of lagged features (which prevents the model from capturing the lead-lag dynamics of Okun's-law-type relationships).

Together, the five plots show that the GWO-XGBoost model relies primarily on RSALES, uses GOVEFF and IPI through bounded relationships shaped by sample structure and feature collinearity, and uses TBAL and UNEMP only marginally. These patterns describe the model's internal decision logic within the predictive-contribution boundary.

4.3 Mapping SHAP findings to the sustainable development goals

The feature attribution results from the SHAP analysis can be linked to specific Sustainable Development Goal targets, providing a structured bridge between the model's predictive findings and the SDG policy agenda. As established earlier in this section, SHAP describes how the model uses its features to generate predictions, rather than the causal economic effects of those features on actual GDP. The mappings, therefore, imply data-driven hypotheses.

The dominant role of consumer demand in the model's predictive structure aligns with the goal of decent work and economic growth, and in particular with the targets of sustaining per capita economic growth in line with national circumstances and achieving higher economic productivity through diversification and technological upgrading. Within the model's learned feature space, fiscal and monetary instruments that operate through household consumption targeted transfers, employment support, and demand-side stimulus act on the channel the model identifies as most predictive. This suggests that consumption-oriented policy levers should be prioritized. The Trade Balance contribution connects to the goal of strengthening global partnerships for sustainable development, and in particular to the target of significantly increasing the export volumes of developing countries. The model assigns a meaningful but secondary predictive weight to the net-exports channel, supporting the case for integrating trade-facilitation measures with domestic-demand instruments rather than treating them as separate policy domains.

4.4 Policy and managerial implications

The empirical findings and SHAP-based attributions translate into a set of practical implications for three stakeholder groups. These are data-driven hypotheses derived from the model's learned feature structure, suitable for guiding decision-making.

4.4.1 Implications for policymakers

Consumer demand emerges as the dominant feature in the model's predictive structure across the full sample window. Within this framework, fiscal and monetary instruments that operate through household consumption targeted transfers, employment support, and demand-side stimulus act on the channel that the model identifies as most predictively informative. The Trade Balance ranks third in the SHAP hierarchy, suggesting that the net-exports channel carries meaningful predictive weight and supporting the integration of trade-facilitation measures with broader macroeconomic stabilization strategy. The relatively small SHAP contribution of unemployment suggests that the contemporaneous unemployment rate may function as a lagging indicator within the model's learned decision structure, and that consumption-based indicators may offer more timely signals for real-time economic monitoring.

4.4.2 Implications for financial analysts

The GWO-XGBoost-SHAP framework can function as a transparent GDP monitoring tool that decomposes predicted GDP movements into feature-level contributions as new monthly data are released. This decomposition reveals whether a predicted change is driven by consumer demand, industrial production, or the external sector, each carrying different implications for sector allocation and risk assessment. The feature ablation results identify retail sales and government effectiveness as the features carrying the most non-redundant predictive information, which may inform the construction of GDP-sensitive factor exposures.

5 Conclusion

This study developed a hybrid forecasting framework that combines GWO and XGBoost with SHAP-based explainability to predict monthly U.S. GDP using macroeconomic indicators enriched with an institutional quality variable. On the test dataset, the proposed GWO-XGBoost model achieved an R^2 of 0.991, with low forecast errors, thereby outperforming the competing machine learning models considered in this study. SHAP analysis showed that Consumer Demand and Government Effectiveness occupy the most influential positions in the model's learned predictive structure, while Trade Balance, Industrial Production, and Unemployment contribute at more moderate levels. When interpreted within the proper predictive contribution framework, these findings generate policy-relevant hypotheses concerning fiscal targeting, governance reform, and trade facilitation.

Several methodological limitations should nevertheless be acknowledged. The empirical analysis is based on a single country and a 25 year monthly sample spanning 1996 to 2020, which may constrain generalizability across structural breaks, cross-country settings, and post-pandemic regimes. Variables were converted to monthly frequency using interpolation in EViews, a procedure that preserves annual anchor points but cannot fully recover intra-period institutional or macroeconomic variation. Model validation

relied on a single shuffled 80:20 holdout split rather than a time series-specific walk-forward design. Although GWO was selected on the basis of its documented effectiveness in the metaheuristic literature and in related forecasting applications, the study did not perform a direct benchmark against several contemporary optimizers. Likewise, dedicated GWO-based tuning was applied only to XGBoost, whereas the benchmark learners were implemented using established parameter settings drawn from prior studies. The SHAP results describe how the model uses the predictors to generate GDP forecasts and should not be interpreted as evidence of causal economic effects. The framework also treats GDP prediction as a prediction task, without lagged predictors or sequence-aware architectures, and therefore does not capture dynamic lead-lag relationships. In addition, the reported metrics do not include several well-known machine learning metrics.

Future research will extend this framework in several directions. The most immediate priority is the adoption of walk-forward temporal validation via an expanding-window design with periodic GWO reoptimization over the test dataset, complemented by directional accuracy measures and regime-conditioned evaluation. A broader metaheuristic benchmark will compare GWO with other contemporary optimizers. A more balanced benchmarking design will also apply comparable optimization pipelines to all baseline learners under a matched search environment and validation protocols. The treatment of temporal dependence can be strengthened through lagged feature engineering, hybrid autoregressive machine learning frameworks such as ARIMA and SARIMA with XGBoost, and sequence-aware architectures including LSTM and Convolutional Neural Networks. Reporting quality will be improved by incorporating more metrics. Broader empirical validation should examine cross-country generalizability, especially across emerging markets and commodity-dependent economies. Finally, complementary causal inference strategies will be implemented to provide a stronger basis for evaluating the policy hypotheses suggested by the SHAP analysis before translating them into substantive recommendations.

Author contributions
O.R.O: Conceptualization, Supervision, Resources, Editing, Methodology, Formal Analysis, Original Draft, D.K: Supervision, Resources, Editing,

Funding
Authors declare no funding was received.

Data availability
The data obtained through the experiments are available upon request from corresponding author.

Declarations

Ethics approval and consent to participate
This study did not involve human participants, animals, or data requiring ethical review.

Consent for publication
Not applicable.

Informed consent
No human subjects were involved, and as such, no informed consent was required for this study.

Competing interests
The authors declare no competing interests.

Received: 4 February 2026 / Accepted: 20 May 2026
Published online: 26 July 2026

References

1. Long LW, Rabab B, Jamil MN. Government spending, political dynamics, and economic recovery: a cross-national analysis of fiscal resilience. Res Glob. Dec. 2025;11:100297. https://doi.org/10.1016/j.resglo.2025.100297.
2. Oancea B, Simionescu M. Jan., 'Gross Domestic Product Forecasting: Harnessing Machine Learning for Accurate Economic Predictions in a Univariate Setting', Electronics, 13, 24, p. 4918, 2024, https://doi.org/10.3390/electronics13244918
3. Abdelghafar S, Darwish A, Ali A. 'Short-Term Forecasting of GDP Growth for the Petroleum Exporting Countries Based on ARIMA Model', in The 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023), March 5–7, 2023, A. E. Hassanien, A. Haqiq, A. T. Azar, K. Santosh, M. A. Jabbar, A. Słowik, and P. Subashini, Eds., Cham: Springer Nature Switzerland, 2023, pp. 399–406. https://doi.org/10.1007/978-3-031-27762-7_37
4. Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res. Oct. 2018;270(2):654–69. https://doi.org/10.1016/j.ejor.2017.11.054.
5. Hopp D. 'Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM)', J. Off. Stat., vol. 38, no. 3, pp. 847–873, Sep. 2022, https://doi.org/10.2478/jos-2022-0037
6. Choi H, Varian H. 'Predicting the Present with Google Trends', Econ. Rec., vol. 88, no. s1, pp. 2–9, 2012, https://doi.org/10.1111/j.1475-4932.2012.00809.x
7. Anesti N, Kalamara E, Kapetanios G. Forecasting with Machine Learning methods and multiple large datasets. Econom Stat Sep. 2024. https://doi.org/10.1016/j.ecosta.2024.08.003.
8. Hopp D. Benchmarking econometric and machine learning methodologies in nowcasting GDP. Empir Econ. May 2024;66(5):2191–247. https://doi.org/10.1007/s00181-023-02515-6.
9. Siami-Namini S, Namin AS. 'Forecasting Economics and Financial Time Series: ARIMA vs. LSTM', Mar. 16, 2018, arXiv: arXiv:1803.06386. https://doi.org/10.48550/arXiv.1803.06386
10. Kim K. 'Financial time series forecasting using support vector machines', Neurocomputing, vol. 55, no. 1, pp. 307–319, Sep. 2003, https://doi.org/10.1016/S0925-2312(03)00372-2
11. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, Burton J, Tutorial 'A. Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications'. Clin Transl Sci. 2025;18(3):e70172. https://doi.org/10.1111/cts.70172.
12. 'Home | Worldwide Governance Indicators', World Bank. Accessed: Nov. 21, 2025. [Online]. Available: https://www.worldbank.org/en/publication/worldwide-governance-indicators
13. Bedane B, Alam MR, Kitenge E. Government Effectiveness and Economic Growth. Econ Bull. 2017;37:222–7.
14. Sori GK, Birhan AT, Thomas ND. 'Effect of Government Effectiveness on Economic Growth: Evidence from Sub-Saharan Africa', Am. J. Econ. Bus. Innov., vol. 3, no. 3, pp. 76–83, Sep. 2024, https://doi.org/10.54536/ajebi.v3i3.3250
15. Chi T-C, Fan T-H, Ghigliazza RM, Giannone D, Zixuan, Wang. 'Macroeconomic Forecasting and Machine Learning', Oct. 13, 2025, arXiv: arXiv:2510.11008. https://doi.org/10.48550/arXiv.2510.11008
16. Kehinde TO, Adedokun OJ, Kareem MK, Akpan J, Olanrewaju OA. 'STL-ELM: A computationally efficient hybrid approach for predicting high volatility stock market', Intell. Syst. Appl., vol. 27, p. 200564, Sep. 2025, https://doi.org/10.1016/j.iswa.2025.200564
17. Kehinde TO, Adedokun OJ, Joseph A, Kabirat KM, Akano HA, Olanrewaju OA. Helformer: an attention-based deep learning model for cryptocurrency price forecasting. J Big Data. Apr. 2025;12(1):81. https://doi.org/10.1186/s40537-025-01135-4.
18. Mehdary A, Chehri A, Jakimi A, Saadane R. 'Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection', Sensors, vol. 24, no. 4, p. 1230, Jan. 2024, https://doi.org/10.3390/s24041230
19. Sheng L, Wu S, Lv Z. Modified Grey Wolf Optimizer and Application in Parameter Optimization of PI Controller. Appl Sci. Jan. 2025;15(8):4530. https://doi.org/10.3390/app15084530.
20. Feng M, Duan Y, Wang X, Zhang J, Ma L. Carbon price prediction based on decomposition technique and extreme gradient boosting optimized by the grey wolf optimizer algorithm. Sci Rep. Oct. 2023;13(1):18447. https://doi.org/10.1038/s41598-023-45524-2.
21. Kehinde TO, Oyedele AA, Kareem MK, Akpan J, Olanrewaju OA. 'Explainable DEA–ensemble approach with golden jackal optimization: efficiency evaluation and prediction for United States information technology firms', Mach. Learn. Appl., vol. 23, p. 100798, Mar. 2026, https://doi.org/10.1016/j.mlwa.2025.100798
22. Elshewey AM, Osman AM. Orthopedic disease classification based on breadth-first search algorithm. Sci Rep. Oct. 2024;14(1):23368. https://doi.org/10.1038/s41598-024-73559-6.
23. Al-Tashi Q, Rais HM, Abdulkadir SJ, Mirjalili S. 'Feature Selection Based on Grey Wolf Optimizer for Oil & Gas Reservoir Classification', in 2020 International Conference on Computational Intelligence (ICCI), Oct. 2020, pp. 211–216. https://doi.org/10.1109/ICCI51257.2020.9247827
24. Yu K, Liang JJ, Qu BY, Chen X, Wang H. Parameters identification of photovoltaic models using an improved JAYA optimization algorithm. Energy Convers Manag. Oct. 2017;150:742–53. https://doi.org/10.1016/j.enconman.2017.08.063.
25. Asghari M, Jahangiri M, Ashrafi K. Machine learning-driven building energy optimization: integrating EnsembleXGBoost sensitivity analysis with meta-heuristic algorithms and carbon economics. Energy Convers Manag X. May 2026;30:101619. https://doi.org/10.1016/j.ecmx.2026.101619.
26. Konyalıoğlu AK, Apaydın TB, Turhan İ, Soydal A, Özcan T. In: Durakbasa NM, Gençyılmaz MG, editors. An Extreme Gradient Boosting Model Optimized with Genetic Algorithm for Sales Forecasting of Retail Stores', in Industrial Engineering in the Industry 4.0 Era. Cham: Springer Nature Switzerland; 2024. pp. 59–67. https://doi.org/10.1007/978-3-031-53991-6_5.
27. Sun W, Wang Y, Zhang L, Chen XH, Hoang YH. Enhancing economic cycle forecasting based on interpretable machine learning and news narrative sentiment. Technol Forecast Soc Change. Jun. 2025;215:124094. https://doi.org/10.1016/j.techfore.2025.124094.
28. Aisy RR, Zulfa L, Rahim Y, Ahsan M. Residual XGBoost regression—Based individual moving range control chart for Gross Domestic Product growth monitoring. PLoS ONE. May 2025;20(5):e0321660. https://doi.org/10.1371/journal.pone.0321660.
29. Silva TC, Wilhelm PVB, Amancio DR. 'Machine learning and economic forecasting: the role of international trade networks', Apr. 11, 2024, arXiv: arXiv:2404.08712. https://doi.org/10.48550/arXiv.2404.08712
30. Dong Thi Ngoc L, Hoan ND, Nguyen H-N. Jan., 'Gross Domestic Product Forecasting Using Deep Learning Models with a Phase-Adaptive Attention Mechanism', Electronics, 14, 11, p. 2132, 2025, https://doi.org/10.3390/electronics14112132

31. Passarella R, Setiawan M, Yamani Z. 'Comparative Analysis of Machine Learning Models for Predicting Indonesia's GDP Growth', Acadlore Trans. AI Mach. Learn., vol. 4, pp. 157–173, Jul. 2025, https://doi.org/10.56578/ataiml040302
32. Soybilgen B, Yazgan E. Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors. Comput Econ. Jan. 2021;57(1):387–417. https://doi.org/10.1007/s10614-020-10083-5.
33. Xia Y. 'A novel combination machine learning model for regional GDP prediction: evidence from China', presented at the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023), Atlantis Press, Oct. 2023, pp. 985–993. https://doi.org/10.2991/978-94-6463-262-0_101
34. Şahin EE. 'Explaining Economic and Environmental Outcomes Through Renewable Energy Quality: An Interpretable Machine Learning Approach Using XGBoost and SHAP', J. Sustain. Dev. Issues, vol. 3, no. 1, pp. 32–52, Jun. 2025, https://doi.org/10.62433/josdi.v3i1.46
35. Baydar MB, Mete M. 'Explaining Logistics Performance, Economic Growth, and Carbon Emissions Through Machine Learning and SHAP Interpretability', Sustainability, vol. 18, no. 2, p. 585, Jan. 2026, https://doi.org/10.3390/su18020585
36. Ali M, Hussein M. Evaluation and prediction of design-time product structural analysis assistance using XGBoost and Grey Wolf Optimizer. Asian J Civ Eng. Apr. 2024;25(3):2415–29. https://doi.org/10.1007/s42107-023-00916-7.
37. Duan Y, Mao Y, Guo Y, Wang X, Gao S. 'COVID-19 Propagation Prediction Model Using Improved Grey Wolf Optimization Algorithms in Combination with XGBoost and Bagging-Integrated Learning', Math. Probl. Eng., vol. 2022, no. 1, p. 1314459, 2022. https://doi.org/10.1155/2022/1314459
38. Oancea B, Forecasting 'AdvancingGDP. The Potential of Machine Learning Techniques in Economic Predictions', in KNOWCON 2024: Knowledge on Economics and Management. Conference Proceedings, 2024. https://doi.org/10.5507/ff.24.24465524
39. Mourtas SD, Drakonakis E, Bragoudakis Z, Mourtas SD, Drakonakis E, Bragoudakis Z. 'Forecasting the gross domestic product using a weight direct determination neural network', AIMS Math., 8, 10, pp. 24254–73, 2023, https://doi.org/10.3934/math.20231237
40. Jallow H, Gibba A, Mwangi RW, Imboga H. GDP prediction of The Gambia using generative adversarial networks. Front Artif Intell. Mar. 2025;8. https://doi.org/10.3389/frai.2025.1546398.
41. Jallow H, Mwangi RW, Gibba A, Imboga H. Transfer learning for predicting of gross domestic product growth based on remittance inflows using RNN-LSTM hybrid model: a case study of The Gambia. Front Artif Intell. Feb. 2025;8. https://doi.org/10.3389/frai.2025.1510341.
42. Chen J, Wu J. The prediction of Chongqing's GDP based on the LASSO method and chaotic whale group algorithm–back propagation neural network–ARIMA model. Sci Rep. Sep. 2023;13:15002. https://doi.org/10.1038/s41598-023-42258-z.
43. Kamruzzaman M, Mondal RS, Islam MK, Rahaman MA, Saha S. Int J Comput Exp Sci Eng. Dec. 2024;10(4). https://doi.org/10.22399/ijcesen.3612. 'AI-Driven Predictive Modelling of US Economic Growth Using Big Data and Explainable Machine Learning'.
44. Akepanidtaworn K, Akepanidtaworn K, 'GDP Nowcasting Performance of Traditional Econometric Models vs Machine-Learning Algorithms. : Simulation and Case Studies', IMF Work. Pap., vol. 2025, no. 252, p. 1, Dec. 2025. https://doi.org/10.5089/9798229033626.001
45. Raghavendran CV, Mouli KC, Latha SB, Motagi AU. A lightweight neural network approach for predicting national Gross Domestic Product (LightNet-GDP) with regression benchmarks. Sci Rep. Jan. 2026;16(1):6634. https://doi.org/10.1038/s41598-026-37672-y.
46. Hisam M, 'GDP Forecasting for Policy: Evaluating ARIMA, Exponential Smoothing, and XGBoost Models for Fiscal, Monetary, and, Planning'W et al. J. Stat. Appl. Probab., vol. 15, no. 2, pp. 185–201, Mar. 2026, https://doi.org/10.18576/jsap/150203
47. Mirjalili S, Mirjalili SM, Lewis A. 'Grey Wolf Optimizer', Adv. Eng. Softw., vol. 69, pp. 46–61, Mar. 2014, https://doi.org/10.1016/j.advengsoft.2013.12.007
48. Yu M, Xu J, Liang W, Qiu Y, Bao S, Tang L. 'Improved multi-strategy adaptive Grey Wolf Optimization for practical engineering applications and high-dimensional problem solving', Artif. Intell. Rev., vol. 57, no. 10, p. 277, Sep. 2024, https://doi.org/10.1007/s10462-024-10821-3
49. Chen T, Guestrin C. 'XGBoost: A Scalable Tree Boosting System', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785
50. Luo S, Wang B, Gao Q, Wang Y, Pang X. 'Stacking integration algorithm based on CNN-BiLSTM-Attention with XGBoost for short-term electricity load forecasting', Energy Rep., vol. 12, pp. 2676–2689, Dec. 2024, https://doi.org/10.1016/j.egyr.2024.08.078
51. Utkarsh, Jain PK. Predicting bentonite swelling pressure: optimized XGBoost versus neural networks. Sci Rep. Jul. 2024;14(1):17533. https://doi.org/10.1038/s41598-024-68038-x.
52. Alsulamy S. 'Predicting construction delay risks in Saudi Arabian projects: A comparative analysis of CatBoost, XGBoost, and LGBM', Expert Syst. Appl., vol. 268, p. 126268, Apr. 2025, https://doi.org/10.1016/j.eswa.2024.126268
53. Kangalli Uyar SG, Ozbay BK, Dal B. Interpretable building energy performance prediction using XGBoost Quantile Regression. Energy Build. Aug. 2025;340:115815. https://doi.org/10.1016/j.enbuild.2025.115815.
54. Federal Reserve Economic Data | FRED | St. Louis Fed. Accessed: Nov. 20. 2025. https://fred.stlouisfed.org/
55. World Development Indicators | DataBank. Accessed: Oct. 20. 2025. https://databank.worldbank.org/source/world-development-indicators
56. Lundberg SM, Lee S-I. 'A unified approach to interpreting model predictions', in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.
57. Peter M, Mirau S, Sinkwembe E, Kasumo C, Guambe C. 'A stacking model integrating GARCH and LSTM with feature interactions for time series volatility prediction', Array, vol. 29, p. 100700, Mar. 2026, https://doi.org/10.1016/j.array.2026.100700

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.