



The consumer price index prediction using machine learning approaches: Evidence from the United States

Tien-Thinh Nguyen^a, Hong-Giang Nguyen^b, Jen-Yao Lee^a, Yu-Lin Wang^c, Chien-Shu Tsai^{d,*}

^a Department of International Business, National Kaohsiung University of Science and Technology, Kaohsiung City, 807618, Taiwan

^b Department of Academic and Students' Affairs, Hue University, Hue City, 49000, Viet Nam

^c Department of Economics, National Chung Cheng University, Chiayi County, 621301, Taiwan

^d Institute of Marine Affairs and Business Management, National Kaohsiung University of Science and Technology, Kaohsiung City, 811213, Taiwan

ARTICLE INFO

Keywords:

Consumer price index
Forecasting
Multivariate linear regression model
Support vector regression model
Autoregressive distributed lag model
Multivariate adaptive regression splines model

ABSTRACT

The consumer price index (CPI) is one of the most important macroeconomic indicators for determining inflation, and accurate predictions of CPI changes are important for a country's economic development. This study uses multivariate linear regression (MLR), support vector regression (SVR), autoregressive distributed lag (ARDL), and multivariate adaptive regression splines (MARS) to predict the CPI of the United States. Data from January 2017 to February 2022 were randomly selected and divided into two stages: 80 % for training and 20% for testing. The US CPI was modeled for the observed period and relied on a mix of elements, including crude oil price, world gold price, and federal fund effective rate. Evaluation metrics—mean absolute percentage value, mean absolute error, root mean square error, R-squared, and correlation of determination—were employed to estimate forecasted values. The MLR, SVR, ARDL, and MARS models attained high accuracy parameters, while the MARS algorithm generated higher accuracy in US CPI forecasts than the others in the testing phase. These outputs could support the US government in overseeing economic policies, sectors, and social security, thereby boosting national economic development.

1. Introduction

The consumer price index (CPI) plays a substantial role in a country's economic price system. It depicts the change in the price of a household's purchase of consumer commodities and services by measuring the relative prices of a collection of representative consumer commodities and services in a certain period [1–3]. As a result, it is used to track the evolution of living costs over time: when the CPI rises, so does the average price level, and vice versa [4]. This variation in CPI results in inflation or deflation, leading to an economic breakdown. Inflation becomes hyperinflation when prices reach uncontrolled levels [5]; moreover, deflation originates from a fall in the CPI general price level due to a decrease in aggregate demand, which leads to economic depression and unemployment. Ultimately, the CPI will be the basis for governments to adjust prices and stabilize the market to avoid inflation, allowing governments

* Corresponding author. Institute of Marine Affairs and Business Management, National Kaohsiung University of Science and Technology, Taiwan. No. 142, Haijhuann Rd., Nanzih Dist., Kaohsiung City 811213, Taiwan.

E-mail addresses: ngtienthinh101096@gmail.com (T.-T. Nguyen), giangnh@hueuni.edu.vn (H.-G. Nguyen), itjylee@nkust.edu.tw (J.-Y. Lee), ecdylw@ccu.edu.tw (Y.-L. Wang), cstsai@nkust.edu.tw (C.-S. Tsai).

<https://doi.org/10.1016/j.heliyon.2023.e20730>

Received 27 August 2023; Received in revised form 3 October 2023; Accepted 4 October 2023

Available online 5 October 2023

2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

to promptly adjust decisions on welfare, social security, and salary increases [6].

The CPI details the prices of various goods and services in a nation to monitor inflation; however, in light of the current global economic volatility, the economies of several supposedly powerful countries are now threatened by excessive inflation. Therefore, accurately predicting the variation in the CPI is crucial to various parts of economics, including fiscal policy, productivity, and financial markets. More importantly, constructing an accurate national CPI predictive model greatly benefits the public and government. This study aims to compare classical econometric and machine-learning models for CPI prediction in the United States.

This study attempts to leverage appropriate machine learning and classical econometric models for forecasting the US CPI from January 2017 to February 2022; hence, support vector regression (SVR), multivariate adaptive regression splines (MARS), multivariate linear regression (MLR), as well as autoregressive distributed lag (ARDL) algorithms were deployed to compare predictive outputs, while several main accuracy indicators, such as the root mean square error (RMSE), mean absolute percentage error (MAPE), R-squared (R^2), and mean absolute error (MAE) were sequentially employed to estimate the accuracy of the forecast algorithms.

An algorithm (Fig. 1) was proposed to examine the US CPI by using MLR, ARDL, MARS, and SVR models to analyze data. First, the database was preprocessed, tested by a statistical approach, and divided into training and testing phases. Subsequently, the MLR, ARDL, MARS, and SVR algorithms were employed to examine the training data and obtain the most accurate indicators. Finally, the performances of the four models were evaluated using metrics for estimating accuracy parameters, and the best estimate procedure for the study was found.

The main contribution of this study is the evaluation of two multivariate (MLR and ARDL) classical econometric models and two machine learning models (SVR and MARS) for US CPI prediction. In addition, these approaches enable a comprehensive analysis and comparison between classical econometric and machine learning models, helping readers understand which models are best suited for US CPI prediction within the same context. It was found that the MLR, SVR, ARDL, and MARS models accurately forecasted the US CPI and could be important tools for policymakers, asset managers, and market participants missing component-specific price forecasts critical to the decision-making process. Table 1 summarizes the definitions of the abbreviations used in this study.

This study's evaluations suggest that the MARS, SVR, ARDL, and MLR models achieve high-accuracy parameters and forecast CPI more accurately than many current baselines. Moreover, the techniques and findings of this study provide policymakers and marketmakers with new insights into projecting price fluctuations from different perspectives. Moreover, the findings of this study could assist the US government in capturing the volatility of the CPI, allowing to minimize the national economy's risks and impose fiscal policy to maintain national economic development.

The remainder of this paper is organized as follows. Section 1 introduces the study, and Section 2 exhibits the related works. Section 3 describes the materials and methods of the MLR, SVR, ARDL, and MARS algorithms. Sections 4 and 5 delineate the results and discussion, respectively, and Section 6 presents the conclusions.

2. Related works

Many previous studies have focused on the factors affecting the CPI. More specifically, oil price, gold price, and the federal fund

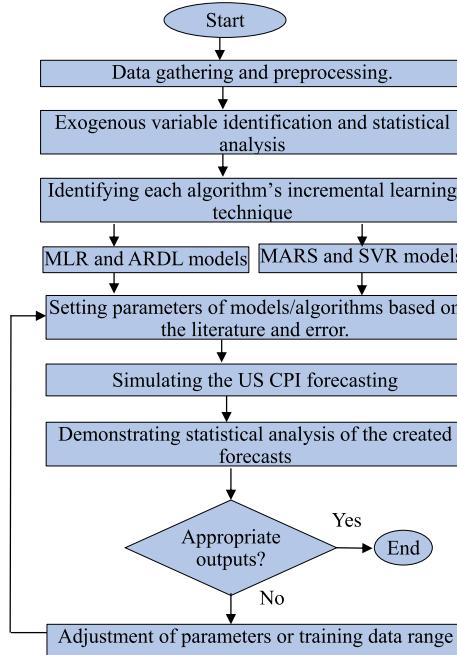


Fig. 1. First experimental steps of the study.

Table 1
Abbreviations meaning.

| Abbreviations | Meaning |
|----------------|--|
| CPI | Consumer Price Index |
| SVR | Support Vector Regression |
| MARS | Multivariate Adaptive Regression Splines |
| MLR | Multivariate Linear Regression |
| ARDL | Autoregressive Distributed Lag |
| RMSE | Root Mean Square Error |
| MAPE | Mean Absolute Percentage Error |
| R ² | R-squared |
| MAE | Mean Absolute Error |
| SD | Standard Deviation |
| SHAP | Shapley Additive Explanation Method |
| FFER | Federal Fund Effective Rate |
| GP | World Gold Price |
| OP | US Crude Oil First Purchase Price |
| KW | Kruskal–Wallis |

effective rate have substantial effects on the US CPI [7–9]. Global oil prices have a significant long-term impact on the CPI [10–12], and the surge in international oil prices positively correlated with the CPI in China from October 1999 to October 2008 [13]. Barakat, Elgazzar, and Hanafy [14] showed a causal relationship between the CPI, interest rate, and exchange rate in Egypt from January 1998 to January 2014. Furthermore, the federal funds rate (FFR) substantially impacts the CPI [15–17], as do global gold prices, but it is also an important predictor of it [18–20]. These studies suggest that oil price, gold price, and the federal fund effective rate can be considered essential elements affecting the US CPI; hence, these factors can be used to model and analyze the prediction of US CPI accuracy using machine learning approaches.

Because the CPI is considered an important macroeconomic factor for measuring inflation, machine learning models can be effective and beneficial for forecasting it. Several studies have employed advanced prediction models to address environmental and energy problems [21–25]. Nonetheless, selecting suitable machine learning models is primarily based on the data structure, with an enormous body of research employing renowned classical econometric and machine learning algorithms, such as MLR and MARS, for prediction [26–32]. Riofrío et al. [33] point out that the SVR technique best forecasts the Ecuadorian CPI, whereas Gjika, Puka, and Zaçaj [34] employed seasonal autoregressive integrated moving average (SARIMA) and artificial neural network (ANN) algorithms to predict the US CPI, covering the period from January 1980 to December 2013. Lidiema [35] deployed the Holt–Winters triple exponential smoothing and SARIMA models to predict the CPI in Kenya from November 2011 to October 2016, while Ülke, Sahin, and Subasi [36] used four machine learning algorithms, including the k-nearest neighbor (kNN), ANNs, SVR model, and an ARDL model to predict US inflation between 1984 and 2014.

Similarly, various studies have deployed regression models to analyze economic growth and environmental quality [37–43]. Kalaycı, Özmen, and Weber [44] used MARS to explicitly analyze the relationship between the CPI and sentiment in the USA from 2001 to 2012. MARS approaches were used to reveal the varying levels of sensitivity of the stock index to macroeconomic and psychological factors from January 2012 to October 2020 [45]. Likewise, Chang [46] employed MARS models to examine the determinants and their impacts on international air passenger flow among nation pairs in 20 countries of the Asia-Pacific Economic Cooperation (APEC) region from 2006 to 2007. MLR was used in the European Union to observe how the CPI, monetary aggregates, and discount rate exchange rates affected inflation [47], while Wang, Wang, and Zhang [48] used SVR to analyze the relationship between the composite index of financial indicators and future inflation in China from 2006 to 2010. On the other hand, Peirano, Kristjanpoller, and Minutolo [49] deployed ANNs, fuzzy inference systems (FISs), artificial neuro-FIS, and SARIMA-ANN as benchmarks to compare the performance of the combined SARIMA-long short-term memory (LSTM) in predicting the inflation rate in five emerging Latin American economies. Likewise, Simionescu [50] employed ANNs and support vector machines (SVMs) to predict

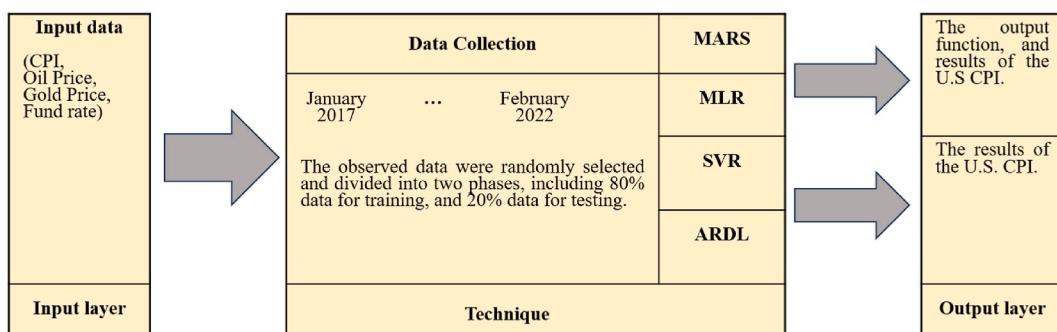


Fig. 2. Successive experimental steps of the study.

inflation in Romania from Q1 2008 to Q4 2021, while Maldeni and Mascarenha [51] used a machine learning method that included random forest (RF), ANNs, extreme gradient boost (XGBoost), SVR, the kNN, and linear regression (LR) for inflation forecast, claiming that it is capable of competing with traditional approaches for prediction accuracy. Barkan et al. [52] applied the novel hierarchical recurrent neural network (HRNN) algorithm to forecast the disaggregated inflation components of the CPI, while Salisu and Isah [53] employed linear time-series algorithms such as the autoregressive integrated moving average (ARIMA) and fractionally integrated versions (ARFIMA) to forecast US inflation from 1957 to 2017. Özmen, Yilmaz, and Weber [54] deployed MARS and LR techniques to predict the natural gas consumption of residential users using daily data from 2009 to 2012. Consequently, these useful outputs offer helpful models and methods for applying the MLR, MARS, SVR, and ARDL algorithms to forecast the US CPI.

3. Data collection and methods

The sequence in Fig. 2 illustrates the steps of this study. Data were collected monthly from January 2017 to February 2022 and comprised 62 observations. The training stage was randomly chosen, contributing 80% of the dataset, and the rest was used for the testing stage. Subsequently, the MLR, SVR, ARDL, and MARS algorithms were applied to examine how the first purchase price of crude oil (OP), world gold price (GP), and federal fund effective rate (FFER) impact the US CPI. Finally, the implementations of the two algorithms are described as base functions and are compared by employing metrics from accuracy estimation parameters such as MAPE, MAE, R², and RMSE.

3.1. Statistical description

The data paradigm searched for different data sources that could be used to build machine-learning models for the US CPI. The US Energy Information Administration is a reliable source for OP data, and GP data were gathered from Statista. Federal Reserve Economic Data were used for the FFER and CPI.

The statistical results in Table 2 show that the standard deviation and mean of the CPI, GP, OP, and FFER are approximately 10.22 and 257.19, 256.32 and 1441.84 US dollars, 13.65 and 55.42 US dollars, and 0.88% and 1.05%, respectively. Additionally, significant differences were observed between the lower 25th, median, and 75th percentiles, indicating strong fluctuations across the four variables. Moreover, the kurtosis and skewness of the independent and dependent variables fluctuate around 0, demonstrating that the data are normally distributed [55].

Pearson's coefficient is numerically equivalent to the correlation coefficient used in linear regression, ranging from -1 to +1. Pearson's correlation coefficient matrix was used to obtain the correlation coefficient between the four variables. Fig. 3 shows that the correlation of OP, GP, and FFER with the CPI equals 0.49, 0.8, and -0.53, respectively, indicating their relatively strong impact on the US CPI. Meanwhile, an incredibly small correlation was observed between the OP, GP, and FFER, except for the correlation between GP and FFER, with a coefficient of -0.78.

The CPI displays statistically significant correlations with other variables. Hence, the coefficient testing result may confirm that CPI is the dependent variable, and the others are independent variables. Table 3 presents the notation and definitions of the variables.

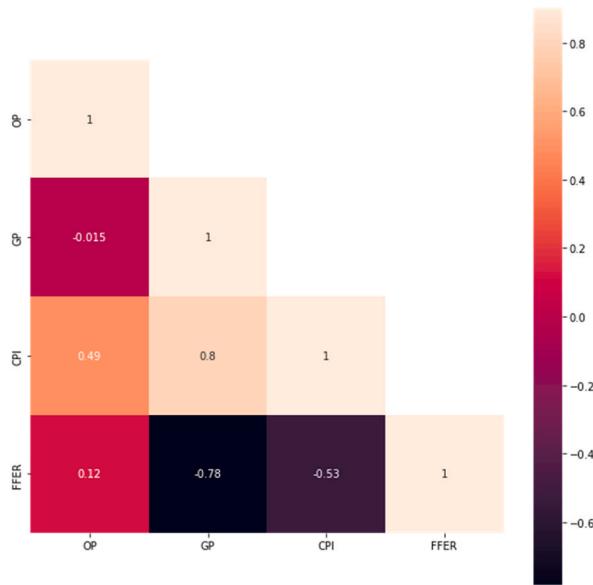
Table 3 summarizes the independent and dependent variables. Furthermore, the availability of data is extremely important for the creation and use of machine-learning models. However, gathering data is not an easy task; therefore, it is crucial to find the best locations for data collection. This presents a potentially useful source for machine-learning modeling.

3.2. Unit root test

The unit root test aims to define whether the data are stationary or not; however, non-stationarity in a time series makes it unpredictable and difficult to model, indicating that the Augmented Dickey–Fuller (ADF) [56] and Phillip–Perron (PP) [57] approaches are regarded as prominent formal tests for determining stationarity [58,59] and are an extension of the DF test that includes an additional lagged term of the dependent variable to remove autocorrelation.

Table 2
Descriptive statistics.

| Statistical Parameter | CPI | GP | OP | FFER |
|--------------------------|------------------|---------------------------|------------|--------|
| Units | Index 1984 = 100 | US dollars per troy ounce | US dollars | % |
| Standard Deviation | 10.219 | 256.324 | 13.648 | 0.879 |
| 1st Quartile | 250.304 | 1281.653 | 47.185 | 0.090 |
| Median | 255.922 | 1441.840 | 55.420 | 1.150 |
| Mean | 257.196 | 1507.968 | 54.442 | 1.056 |
| 3rd Quartile | 260.629 | 1783.173 | 62.800 | 1.828 |
| Min | 243.618 | 1192.100 | 15.180 | 0.050 |
| Max | 284.182 | 1968.630 | 89.410 | 2.420 |
| Kurtosis | 0.319 | -1.586 | 0.992 | -1.500 |
| Skewness | 0.902 | 0.320 | -0.350 | 0.157 |
| Coefficient of variation | 0.004 | 0.17 | 0.25 | 0.82 |

**Fig. 3.** Pearson's correlation coefficient matrix.**Table 3**

List of independent and dependent variables in the dataset.

| Kind | Variable | Definition | Data Source |
|-----------------------|----------|-----------------------------------|--------------------------------------|
| Independent Variables | x_1 | US Crude Oil First Purchase Price | US Energy Information Administration |
| | x_2 | World Gold Price | Statista |
| | x_3 | Federal Fund Effective Rate | Federal Reserve Economic Data |
| Dependent Variable | y | US Consumer Price Index | Federal Reserve Economic Data |

$$\Delta Y_t = C_0 + \alpha_1 Y_{t-1} + \alpha_2 t + \sum_{i=1}^z r_i \Delta Y_{t-i} + \varepsilon_t \quad (1)$$

$$\Delta Y_t = C_0 + \alpha_1 Y_{t-1} + \alpha_2 t + \varepsilon_t \quad (2)$$

where Y represents a time series, t symbolizes a linear time trend, Δ denotes the first difference operator, C_0 is a constant, z is the optimal lag of the ADF equation, and ε represents a random error term in Equations (1) and (2). The hypotheses of the ADF and PP tests are as follows:

H0. The series is not stationary (obtains the unit root).

H1. The series is stationary (no unit root).

H0 is rejected if the p-values of the series of observations are significant at the 1%, 5%, or 10% level.

3.3. The Johansen Co-integration test

The series are integrated because the unit root tests show that they are stationary at the first level; however, even when they are integrated, it does not guarantee that they will behave consistently over time. Cointegration analysis can be used to identify long-term relationships between two nonstationary series, and it can be conducted using certain tests. A Johansen cointegration test is used to examine the long-term relationship between the co-integrated series, indicating that this study can determine the number of co-integration relationships and the parameters deploying the Johansen co-integration test [60]. Furthermore, the trace test examines the null hypothesis of r co-integrating vectors against the alternative hypothesis of w co-integrating vectors, where w symbolizes the number of endogenous variables and r (rank of the data matrix) = 0, 1, 2, ..., $w-1$ [61].

3.4. Multivariate linear regression (MLR)

MLR is a statistical method for predicting the output variable based on a set of explanatory variables, and it is used to define the linear relationship between the independent variable x and the dependent variable y being estimated. It has the key benefit of ease of

use [62]. The fundamental MLR model [63] is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, 3, \dots, n \quad (3)$$

This model can be represented in matrix notation as [64,65].

$$y = X\beta + \varepsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

y_i is the real value response for the i th observation; β_0 is the bias term (regression intercept); $\beta_i (i = 1, \dots, k)$ values are coefficients (the prediction's regression slope); x_{ik} is the k th estimation for the i th observation, and $\varepsilon_i \text{iid } N(0, \sigma^2)$ is the Gaussian error term in Equation (3).

3.5. Multivariate adaptive regression splines (MARS)

Initially introduced by Yildiz and Özdemir [45], the MARS method is considered a versatile approach for modeling additive relationships or interactions with fewer variables, and it is a highly accurate method for modeling systems based on datasets. The computational space is divided into sub-ranges of input observations, and the correlation between input and output observations is identified. This approach can also reveal the connection between dependent and independent variables for each anticipated phenomenon by matching a simple regression to each input value to forecast the output. The MARS method divides the input value space into several units and then matches a spline function on these units while relying on a divide-and-conquer strategy, providing a training dataset in different sections, each with its own regression line [66,67]. The key benefit of the MARS technique is that it highlights input factors that have a significant impact on output parameters; therefore, this approach can be applied to small and large datasets. Moreover, the MARS algorithm uses backward and forward stepwise processes to explain and regulate the intricate nonlinear mapping between input and output variables. The input variable x and new output y are predicted using one of the two fundamental functions and applying a value or knot of variables, indicating the inflection point along the input range [68]. Furthermore, the primary benefit of the MARS model is its capacity to provide precise details regarding the internal procedures performed during the model development process [69]. The overall structure of the MARS method [45] is presented by the following Equation (4):

$$y = f_{(x)} = \beta_0 + \sum_{j=1}^P \alpha_j \beta_j \quad (4)$$

where y denotes the dependent variable forecasted by the function $f_{(x)}$; β_0 is the constant parameter; P denotes the number of terms, each of which is established by a coefficient $\alpha_j, j \in \{1, \dots, P\}$; β_j symbolizes an individual basis function, and x_j is considered a predictor variable. In addition, the fundamental functions $\text{Max}(0, H - x)$ and $\text{Max}(0, x - H)$ are univariate and do not have to each be present if their β coefficients are 0. Notably, H is called "knots" or "hinges," and x denotes the independent variable.

The backward stepwise function eliminates the original functions one at a time until the criterion of "lack of fit" is at its lowest point. To avoid overly weighty equations in the next step, less applicable basis functions are pruned by employing generalized cross-validation (GCV) [70]:

$$GCV = A * \sum_i (y_i - \hat{f}(x)) / N \quad (5)$$

where $A = \left[1 - \frac{C(M)}{N}\right]^{-2}$ as well as $C(M) = (M+1) + dM$ symbolize the complexity function [71], and d implies a penalty for each basis function. In addition, $\hat{f}(x)$ denotes the fitted response value; N is the number of data points, and M is the number of basis functions in Equation (5). The GCV criterion is the average of residual error multiplied by a penalty for modifying the variability associated with more parameter evaluations in the model [72].

3.6. Support vector regression (SVR)

The SVR algorithm is applied to find a suitable hyperplane in higher dimensions to be appropriate for the data with an acceptable threshold error 2ε [73], and it is a technique that has been widely used in regression assignments. Moreover, a hyperplane is a line separating two data classes in a dimension higher than the real dimension; therefore, it is described in SVR as a line that aids in forecasting the target value. Likewise, boundary lines are deployed to generate a margin among the data points and are drawn around the hyperplane at a distance of ε (epsilon). The acceptable threshold error is 2ε , and this value is the distance from the hyperplane to the boundary line (details in Fig. 4) [74,75]. In this study, the value of ε is selected as 0.01. The primary benefit of SVR is that its optimization process relies on the structural risk minimization concept, which seeks to reduce the upper bound of the general error

containing the sum of the training errors [76].

Where $\frac{1}{2}\|w\|^2$ is the confidence interval; and as the application of ε is insensitive, w is the weight vector, where $\|w\|$ denotes the magnitude of the normal vector with respect to the surface ; y_i represents the dependent variables; \hat{y}_i denotes the dependent output variables, x represents the independent input variables; and b implies the threshold value [77]. To simplify the mathematical notation for multidimensional data, x is increased by 1, and b is added to the w vector to achieve multivariate regression [78].

3.7. Autoregressive distributed lag (ARDL)

This study also deploys the ARDL model to identify long-run relationships over the existing traditional cointegration test because this technique is appropriate in a mixed order of integration of variables; whether variables are stationary at a level I (0) or I (1) does not affect the results. A key advantage of the ARDL model is that it can be applied to a small sample [79]. It was developed by Pesaran and Shin [80] and Pesaran, Shin, and Smith [81], and it is defined as follows [82]:

$$\Delta CPI_{t,i} = \alpha_1 + \sum_{i=1}^m \theta_{1i} \Delta CPI_{t-i} + \sum_{i=0}^n \theta_{2i} \Delta GP_{t-i} + \sum_{i=0}^k \theta_{3i} \Delta OP_{t-i} + \sum_{i=0}^r \theta_{4i} \Delta FFER_{t-i} + \pi_1 CPI_{t-1} + \pi_2 GP_{t-1} + \pi_3 OP_{t-1} + \pi_4 FFER_{t-1} + \omega_t \quad (6)$$

where θ symbolizes the short-run coefficients; π is the long-run coefficients; Δ represents the first differences of the variables; and m, n, k, r denotes the lags of the variables in Equation (6).

3.8. Evaluation metrics

This study used four metrics to discuss the performance attainment of the machine-learning models. These accuracy measurement parameter metrics consist of MAPE [83] in Equation (7), MAE [84] in Equation (8), RMSE [85] in Equation (9), and R^2 [86] in Equation (10), which were deployed to estimate the predicted values.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|} \quad (7)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

where n is the total number of collected values; and y_i , \hat{y}_i , and \bar{y} represent the actual value, predicted output, and arithmetic mean of the data estimated at period i , respectively.

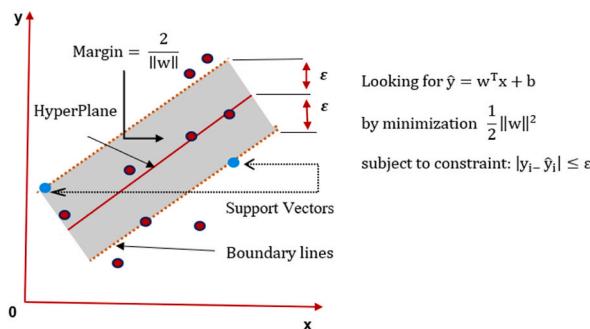


Fig. 4. Graphical representation of the SVR model.

4. Results

4.1. Testing stationary results

The results in [Table 4](#) demonstrate that the time series is not stationary at the chosen level since the p-value for the Augmented Dickey–Fuller and Phillips-Perron tests is higher than 0.01 for all lags. However, at the first difference, all observable variables achieve stationarity, indicating that the time series of these variables is predictable and thus easy to model [87].

4.2. Johansen Co-integration test results

The Johansen co-integration trace test results in [Table 5](#) reject the null hypothesis ([H0](#): no co-integration). Hence, these results also indicate unique cointegrating equations at the 0.05% significance level between variables such as CPI, OP, GP, and FFER. Additionally, these outputs imply a long-run relationship among the observed variables [52].

4.3. Decomposition technique of the CPI, GP, OP, and FFER time series

The CPI predictive graphics aim to define the most appropriate forecasting algorithm. Therefore, [Fig. 5](#) shows that time-series decomposition considers a series as a combination of level, trend, seasonality, and noise components; hence, it also supplies a valuable abstract algorithm for a deeper understanding of problems during time-series examination and prediction. Regarding the trend component, the line graph shows that most variables were stable during this period except for GP. Regarding seasonality, the lines for GP (red line) and OP (purple line) are highly seasonal, while the CPI (blue line) and fund rate (black line) indicate low seasonality from January 2017 to February 2022. The residual component represents the random variation in the time series.

4.4. Output function

The output function of the MLR and MARS of the US CPI is defined as follows:

$$MARS_{CPI} = 359.72 + 0.62F_1 - 0.2F_2 + 7.01F_3 - 0.08F_4 + 0.02F_5 - 0.05 * x_2$$

where $F_1 = \max(0, x_1 - 60.67)$, $F_2 = \max(0, 60.67 - x_1)$, $F_3 = \max(0, x_3 - 1.58)$, $F_4 = \max(0, 1790.43 - x_2)$, and $F_5 = \max(0, x_2 - 1331.3)$.

F_1 , F_2 , F_3 , F_4 , and F_5 are fundamental functions. F_1 is considered the maximum of 0 and $x_1 - 60.67$. A negative sign before the maximum value represents the minimum value.

$$MLR_{CPI} = 183.02 + 0.37 * x_1 + 0.04 * x_2 + 1.12 * x_3$$

in addition, in MLR forecasting, a t-test was conducted to determine the MLR independent variables that were significant [88]; the findings show that OP, GP, and FFER have a positive and significant impact on the US CPI, as shown in [Table 6](#).

4.5. MLR, MARS, ARDL, and SVR algorithm estimation results

The scatter graphs from the different models in [Fig. 6\(a\)](#) and (b) imply that the predicted values generated by the MLR, SVR, ARDL, and MARS models are also moderately close to the actual US CPI. Similarly, the scatter plot for the MARS algorithm is more fitted to the actual line in the testing stage, and the line for the SVR model more closely matches the actual line in the training phase.

When it comes to the testing and training phases (shown in [Tables 7 and 8](#)), the MAE parameters of the two algorithms vary from

Table 4

Unit root test.

| Variable | ADF Statistics | | | PP Statistics | | |
|----------------------------|------------------------|-----------------------------|--|------------------------|-----------------------------|---|
| At Level | | | | | | |
| CPI | None 2.772 (0.998) | Intercept 2.909 (1.000) | Trend and Intercept 1.692 (1.000) | None 4.288 (1.000) | Intercept 3.993 (1.000) | Trend and Intercept 2.946 (1.000) |
| OP | 0.692 (0.862) | -0.903 (0.780) | -1.049 (0.929) | 0.630 (0.849) | -1.032 (0.736) | -1.141 (0.913) |
| GP | 1.116 (0.929) | -0.766 (0.821) | -1.983 (0.599) | 1.498 (0.966) | -0.704 (0.838) | -1.635 (0.767) |
| FFER | -0.811 (0.361) | -0.941 (0.768) | -2.019 (0.579) | -0.709 (0.405) | -0.721 (0.833) | -1.788 (0.698) |
| At First Difference | | | | | | |
| CPI | None -1.068 (0.255) | Intercept -2.159 (0.223) | Trend and Intercept -4.194*** (0.008) | None -1.776* (0.07) | Intercept -2.753* (0.07) | Trend and Intercept -3.482** (0.049) |
| OP | -5.847*** (0.000) | -5.901*** (0.000) | -5.955*** (0.000) | -4.576*** (0.000) | -4.524*** (0.000) | -4.512*** (0.000) |
| GP | -5.828*** (0.000) | -5.992*** (0.000) | -5.941*** (0.000) | -5.771*** (0.000) | -5.903*** (0.000) | -5.848*** (0.000) |
| FFER | -4.801*** (0.000) | -4.771*** (0.000) | -4.933*** (0.000) | -4.709*** (0.000) | -4.676*** (0.000) | -4.841*** (0.000) |

[Table 4](#) represents values of t-statistics and p-values in parentheses.

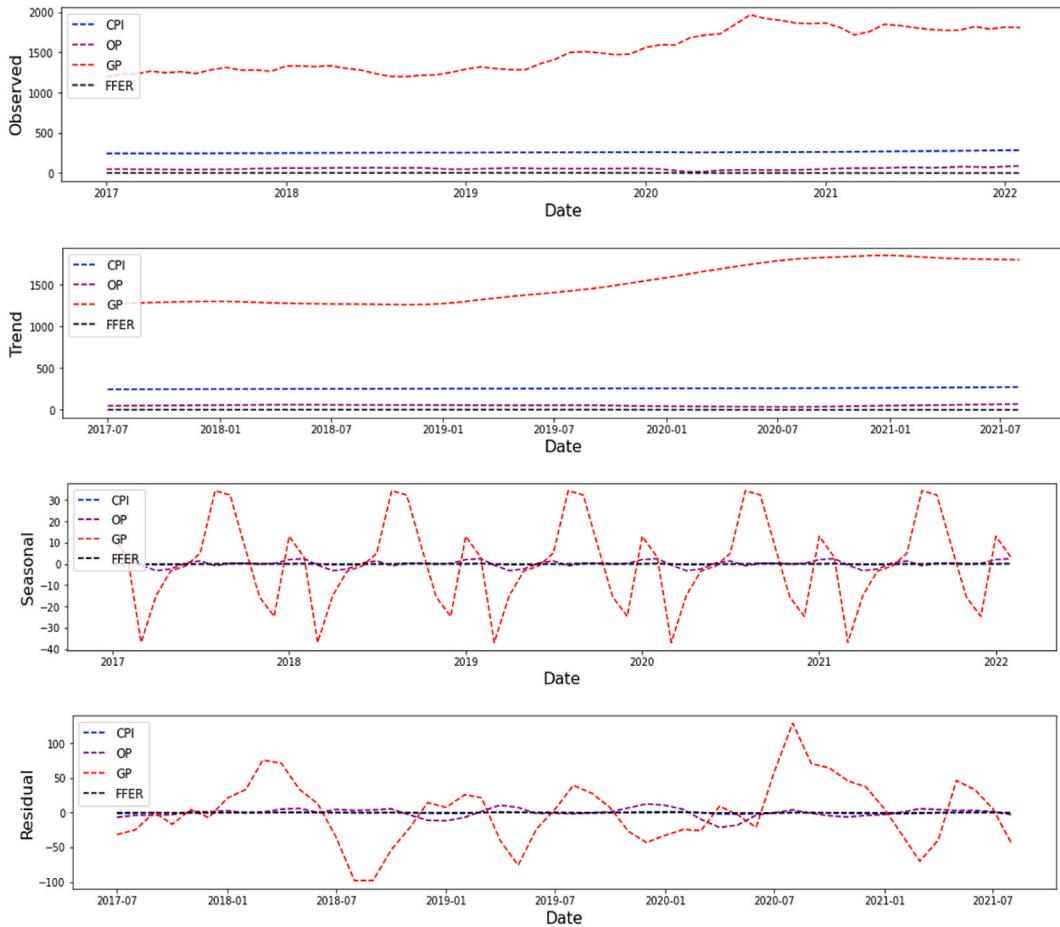
***/**/* imply significance at 1%, 5%, and 10% levels.

None, Intercept, and Trend and Intercept are unit root test conditions.

Table 5

The Johansen co-integration test result with CPI, OP, GP, and FFER.

| Hypothesized No. of CE(s) | Eigenvalue | Trace Statistics | 0.05 Critical Value | Probability |
|---------------------------|------------|------------------|---------------------|-------------|
| None ($r = 0$) | 0.288 | 50.029 | 47.856 | 0.03 |
| At most 1 ($r \leq 1$) | 0.242 | 29.948 | 29.797 | 0.05 |
| At most 2 ($r \leq 2$) | 0.155 | 13.571 | 15.494 | 0.09 |
| At most 3 ($r \leq 3$) | 0.06 | 3.614 | 3.841 | 0.06 |

Note: r is a rank of the data matrix.**Fig. 5.** The calibrating vector decomposition of the time series of the four analyzed variables.**Table 6**

Multiple linear regression results.

| Variable | Coefficient | t-Statistic | Probability |
|--------------|-------------|-------------|-------------|
| C | 183.02 | 33.439 | 0.000 |
| x_1 (OP) | 0.37 | 11.323 | 0.000 |
| x_2 (GP) | 0.035 | 11.483 | 0.000 |
| x_3 (FFER) | 1.12 | 1.348 | 0.049 |

0.1 to 3.2, and the RMSE values range from 0.1 to 3.9. The MAPE values of the testing and training phases for both models fluctuate from 0.001 to 0.013. However, the MAPE value of the MARS algorithm achieves the highest accuracy in predicting the US CPI, with a value of 0.004 for the testing stage. Meanwhile, the SVR model attains the highest accuracy in forecasting the US CPI, with a parameter approximately equal to 0.001 for the training phase. Nevertheless, the MLR model also achieves high-accuracy parameters, and the MAPE values of the MLR model are higher than those of the MARS model by 0.006 for the training phase and 0.009 for the testing

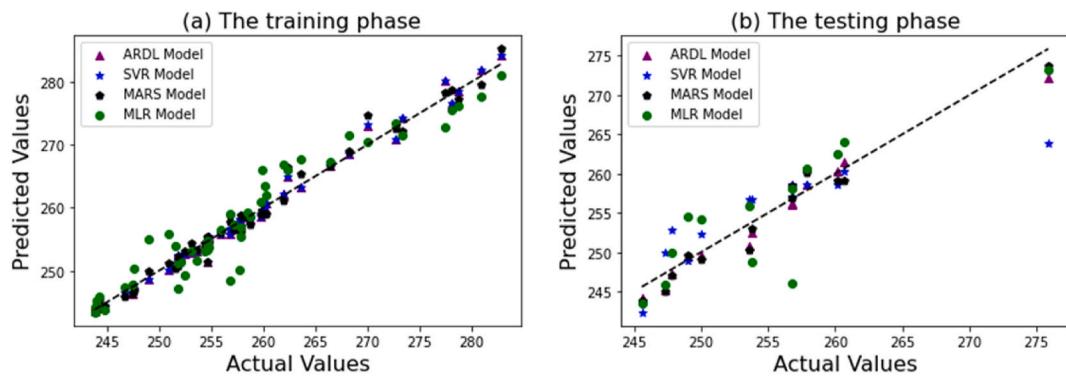


Fig. 6. The actual and forecast CPI based on ARDL, SVR, MLR, and MARS algorithms for (a) the training phase and (b) the testing phase.

Table 7

Accuracy parameters for US CPI forecast during the training period.

| | MLR | MARS | SVR | ARDL |
|--------------------|-------|-------|-------|-------|
| MAPE | 0.009 | 0.004 | 0.001 | 0.002 |
| MAE | 2.400 | 0.961 | 0.100 | 0.735 |
| (Index 1984 = 100) | | | | |
| RMSE | 3.199 | 1.170 | 0.100 | 1.084 |
| (Index 1984 = 100) | | | | |
| R ² | 0.899 | 0.988 | 0.998 | 0.989 |

Table 8

Accuracy parameters for US CPI forecast during the testing period.

| | MLR | MARS | SVR | ARDL |
|--------------------|-------|-------|-------|-------|
| MAPE | 0.013 | 0.004 | 0.009 | 0.005 |
| MAE | 3.219 | 0.933 | 2.400 | 1.29 |
| (Index 1984 = 100) | | | | |
| RMSE | 4.046 | 1.200 | 3.199 | 1.688 |
| (Index 1984 = 100) | | | | |
| R ² | 0.755 | 0.976 | 0.500 | 0.958 |

phase; this implies that the MARS model yields a more accurate forecast than the MLR model. For the testing stage of this study, the MAPE of the MLR model, 0.013 is higher than that of the SVR and MARS models by nearly 0.001 and 0.009, respectively, showing that the MARS model outperforms the SVR and MLR algorithms. More importantly, the ARDL model also attains high accuracy values, and the MLR model's MAPE values are higher than the ARDL model's MAPE parameter by 0.007 for the training phase and 0.008 for the testing phase.

Additionally, the Kruskal–Wallis (KW) test was applied to compare the distribution of the computed and estimated US CPI values indicated in [Tables 9 and 10](#); this is regarded as a nonparametric technique to analyze the variance and examines the null hypothesis by which the predictions among the groups show no significant differences. The parametric equivalent of the KW test is the one-way analysis of variance (ANOVA), assuming the data's normal distribution [89]; as a result, the KW test results for the MLR, MARS, SVR, and ARDL models in both training and testing phases confirm that we cannot reject the null hypothesis at the 5% significance level.

The Taylor diagrams analyze forecast performance and actual values, employing the standard deviation (SD) and correlation used in estimating the models. They allow to compare different model predictions based on their standard deviation and correlation with observations [90], and they have been generally employed to compare the performances of different regression models [91].

Table 9

Results of the KW test with a 0.05 level of significance for the training stage.

| Model | P value | H ⁰ _a |
|-------|---------|-----------------------------|
| MLR | 0.482 | accept |
| MARS | 0.481 | accept |
| SVR | 0.480 | accept |
| ARDL | 0.481 | accept |

Table 10

Results of the KW test with a 0.05 level of significance for the testing stage.

| Model | P value | H_a^0 |
|-------|---------|---------|
| MLR | 0.462 | accept |
| MARS | 0.463 | accept |
| SVR | 0.461 | accept |
| ARDL | 0.471 | accept |

Thereafter, the Taylor chart indicates the standard deviation and correlation between the real and predicted values for the four algorithms and general consistency between observed and predicted values when the correlation value approaches 1, as shown in Fig. 7. The MLR, MARS, ARDL, and SVR models for testing, with SD = 8.172, 7.809, 7.550, and 5.486, respectively, are close to the actual CPI (SD = 7.799). With regard to the standard deviation of the training phase, the SVR model's standard deviation (SD = 10.577) is closest to the actual values (SD = 10.693); consequently, the Taylor diagram's scatter plots prove that the MARS, SVR, ARDL, and MLR models are the four most appropriate methods for forecasting the US CPI.

4.6. Feature analysis with shapley additive explanation method (SHAP)

The Shapley additive explanation (SHAP) method indicates that each feature is given a relevance value by SHAP for a specific prediction. Specifically, a positive SHAP value implies that the input has a positive impact on the output, whereas a negative SHAP value implies that the input has a negative impact on the output [92]. As shown in Figs. 8–11, the SHAP values for the ARDL, SVR, MLR, and MARS models are obtained by ordering the features by the sum of the magnitudes of the SHAP values during both the training and testing phases, while applying the SHAP values to point out the distribution of the impacts of each feature on the technique output. The color represents the feature's value (high red or low blue). More importantly, GP and OP are the two most significant features of the four proposed models, contributing to a positive prediction of the US CPI, while the FFER seems to be a less significant feature for both the training and testing stages. Additionally, high GP, OP, and FFER make a significant positive contribution to US CPI prediction.

5. Discussion

In this study, the MLR, SVR, ARDL, and MARS algorithms were deployed to predict the CPI in the US, relying on monthly data collected from January 2017 to February 2022. For the testing stage, the MLR's MAPE value was the highest, with a coefficient of 3.219, whereas the MARS algorithm had the lowest value, with a coefficient of 0.933, indicating a gap of nearly 2.286 between the two model outputs. Furthermore, the MAPE values for the MLR, ARDL, and MARS models were higher than the MAPE parameter of the SVR model by 0.008, 0.001, and 0.003, respectively, confirming that the SVR model produces a more accurate forecast than the MLR, ARDL, and MARS models during the training phase. Furthermore, the MAPE of the MARS model, 0.004, was lower than that of the MLR, ARDL, and SVR algorithms by 0.027, 0.001, and 0.028, respectively, indicating that the MARS algorithm provides a more accurate forecast than the others for the testing stage.

Regarding research associated with MARS, MLR, ARDL, and SVR models, Sahraei et al. [93] applied the MARS algorithm to forecast transport energy demand, relying on a combination of elements containing gross domestic product (GDP), population, ton-km, vehicle-km, passenger-km, and OP from 1975 to 2019. The third MARS model had the lowest RMSE index of 669.18 and the highest R^2 index of 0.988. Haouraji et al. [94] deployed MLR to predict Morocco's residential energy needs by collecting data monthly from July 2017 to February 2021; these results also showed that the attained RMSE and MAPE parameters for the MLR algorithm were 20.11 and 3.77%, performing well in prediction, being significantly higher by 16.891 and 2.47% for the MLR model, 18.91 and 3.37% for the MARS model, and 16.911 and 2.87% for the SVR in this study's testing phases. Wang, Wang, and Zhang [48] used SVR to analyze future inflation in China from 2006 to 2010, and the results showed that the achieved RMSE value for the SVR model was approximately 0.16, which is considerably higher than the RMSE of this study's SVR model (0.1) in the training stage. Last, Ülke, Sahin, and Subasi [36] demonstrated that the prediction of ARDL and SVR was highly accurate, with RMSE values of 2.66 and 3.11, respectively. Hence, these outputs showed that the k-NN outperformed the ARDL and SVR models, and the RMSE value of the ARDL model was also considerably higher than that of the ARDL model in this study, with an RMSE parameter of 1.69 in the testing phase. To compare advanced prediction models, Zhu et al. [25] applied a back propagation neural network (BPNN) time-series long-term predicting algorithm; the R^2 values of MLR, ARDL, MARS, and SVR achieved higher parameters than the attained R^2 values of the BPNN model in daily temperature time-series prediction horizons: 6, 19, and 165. Regarding the granular computing method, Enke and Mehdiyev [95] applied a hybrid neuro-fuzzy model to forecast the US CPI and obtained an RMSE of 0.829 in their proposed hybrid model, which was significantly higher than the RMSE value of the SVR model in the training stage of this study.

As far as the studies related to CPI prediction are concerned, Ali and Mohamed [87] used autoregressive integrated moving average errors (ARIMAX) to predict the CPI in Puntland State from July 2017 to February 2021. Hence, the MAPE of the ARIMAX model (2.679) was higher than those of the MLR and MARS models by nearly 0.01 in this study's testing and training stages. In addition, Yang and Guo [96] employed a deep learning model with recurrent neural networks and a gated recurrent unit (GRU-RNN) to train and estimate the CPI from April 2005 to June 2021 in China. The MAPE of the proposed model (0.004) exactly matched that of this study's MARS model (0.004) in both the testing and training stages. Kalayci, Özmen, and Weber [44] found that MARS forecasting performed

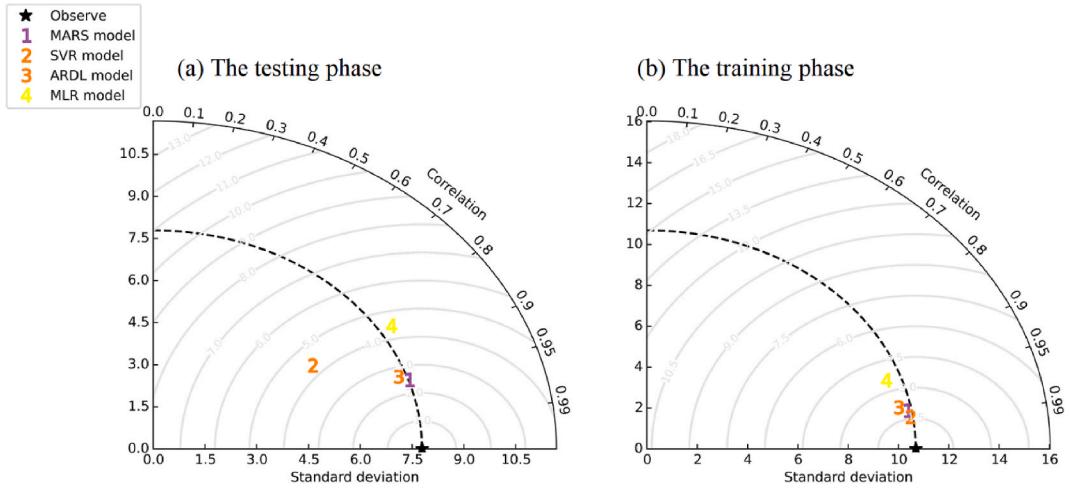


Fig. 7. Taylor diagram demonstrating a statistical comparison of model performance for (a) the testing phase and (b) the training phase.

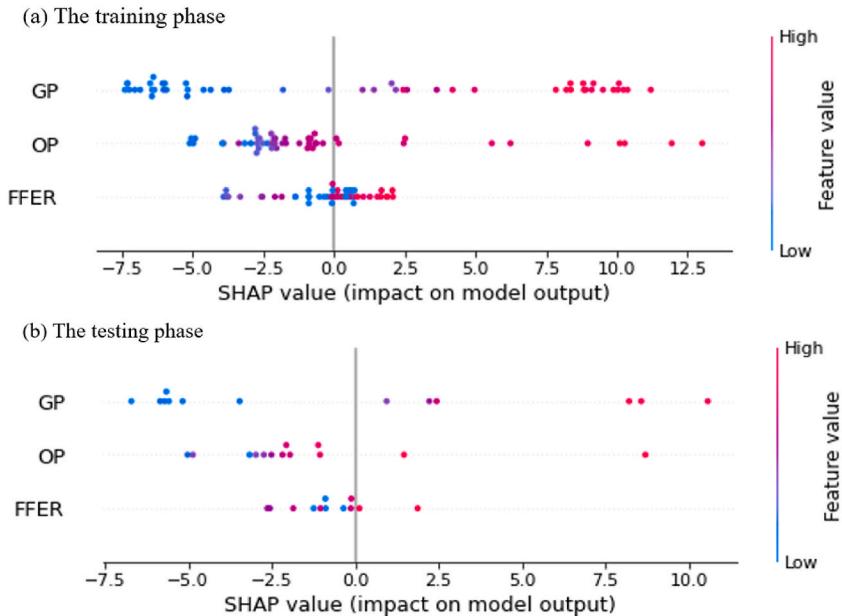


Fig. 8. Feature importance based on the ARDL model for (a) the training phase and (b) the testing phase.

well, with RMSE being close to 1, and obtained the same accuracy parameter as the MARS model used in the training and testing stages of this investigation. Additionally, the study of Groeneveld and Meeden [55] revealed that the MAPE of the SVR and ANN algorithms in predicting inflation in Romania fluctuated between 0.02 and 0.07, indicating that the SVR algorithm outperformed the ANN algorithm. Similarly, these output parameters were much higher than the MLR, SVR, and MARS accuracy values in this study, with MAPE values of 0.01, 0.01, and 0.004, respectively. Riofrío et al. [33] obtained a MAPE of 0.001 in their SVR model, which is significantly close to the MAPE value of the SVR model in the training stage of this study. Although the observed period is associated with significant events that substantially impact US inflation, such as the COVID-19 pandemic and US tariff policies on US imports from China, the SVR, MARS, ARDL, and MLR models proposed in this study still attain more accurate CPI forecasts than most of the studies described above.

Subsequently, economists could use the MLR, SVR, ARDL, and MARS models to predict the CPI in the US and forecast other macroeconomic variables. More importantly, although the MARS, SVR, ARDL, and MLR algorithms are considered classical methods, they still generate excellent predictions of the US CPI. Therefore, these models can be applied to predict the CPI of other nations around the globe.

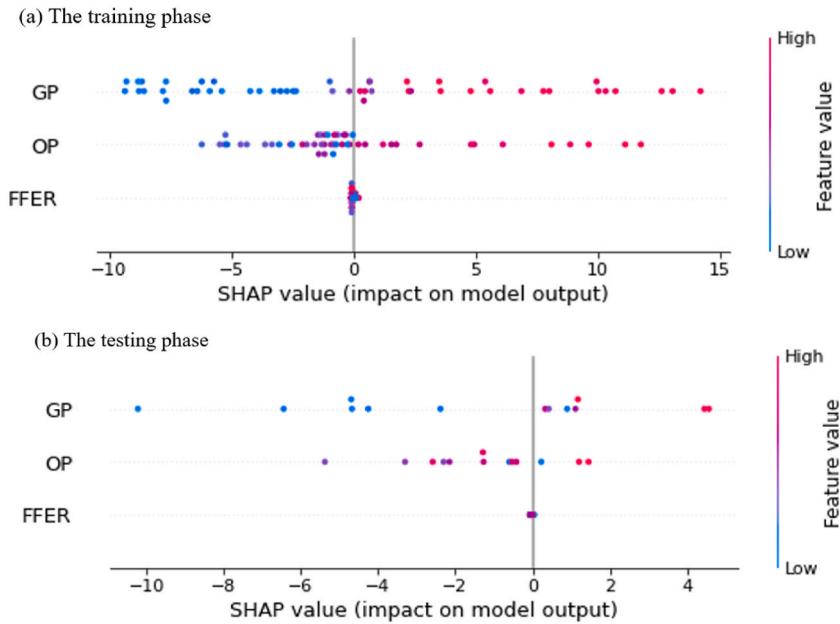


Fig. 9. Feature importance based on the SVR model for (a) the training phase and (b) the testing phase.

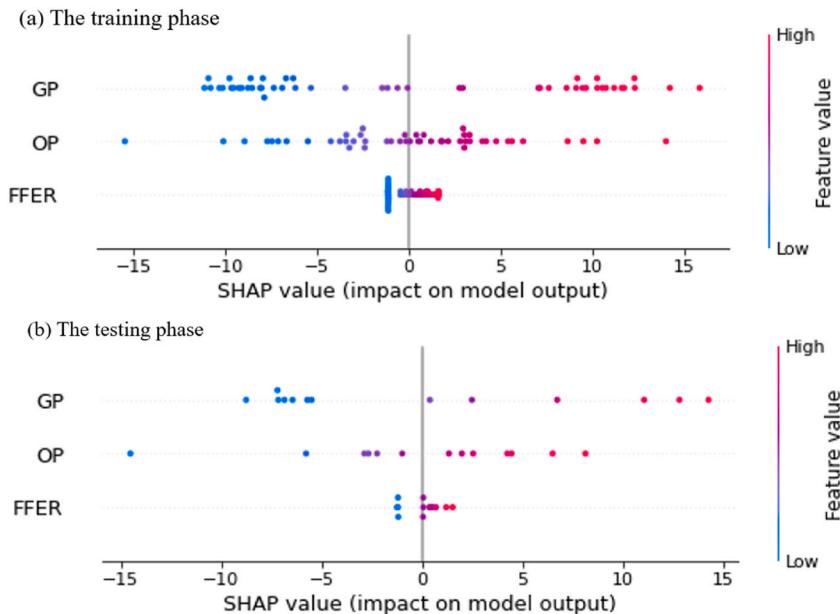


Fig. 10. Feature importance based on the MLR model for (a) the training phase and (b) the testing phase.

6. Conclusion

Although the four proposed models accurately forecasted the US CPI, the MARS algorithm outperformed the MLR, ARDL, and SVR algorithms in the testing phase. Moreover, a strong relationship existed between GP, FFER, and OP in the US during the observed period. More accurate US CPI predictions that rely on these four forecasting techniques have numerous direct implications. By establishing accurate CPI expectations, policymakers can create optimum strategies to lessen the negative consequences of this phenomenon with fewer detrimental effects on the population. Moreover, this study provides investors with a broad picture of changes in the US CPI, allowing suitable investment decisions. Maintaining pricing stability assists in raising living standards and relieving social tensions. Furthermore, machine-learning algorithms can examine historical economic and financial data to predict future price changes. Machine-learning models can process and analyze large volumes of historical financial and economic data, enabling them to

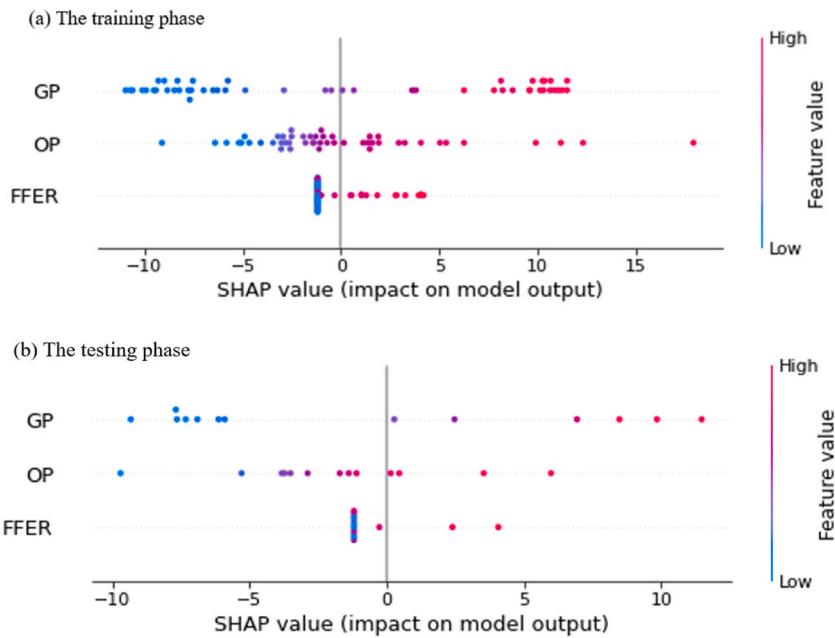


Fig. 11. Feature importance based on the MARS model for (a) the training phase and (b) the testing phase.

identify trends and correlations that might not be apparent through traditional analysis. These predictions can be used to anticipate market trends, asset price fluctuations, and the state of the economy; at the same time, machine-learning and classical econometric models can be deployed to estimate the efficiency of current economic policy by comparing the actual and predicted values, contributing to the fact that policymakers and market makers can evaluate whether their policies attained the desired economic stability and development, and make wise monetary and fiscal policy choices to maintain and expand the national economy. Subsequently, these models can be considered good tools for forecasting, helping market makers and policymakers predict price fluctuations and effectively mitigate national economic risks.

This study had limitations as the number of independent variables did not meet the study's original expectations. Additionally, the predictions were based on the data period, allowing for the potential for concept drift under many circumstances. In machine learning, the term "concept drift" refers to the alteration of relationships between input and output data over time [97]. This may occur when the training algorithm depends primarily on data series with downward tendencies that overpower upward tendencies. Testing is performed on data dominated by upward trends. More importantly, using machine learning algorithms requires substantial time to perform numerous experiments to find the optimal hyperparameters for current and future research. In addition, this study has not fully explored the potential economic factors affecting the US CPI and used small input data that have not achieved the most optimal results from these proposed models.

The CPI is not only considered an important macroeconomic factor in forecasting the possible inflation or deflation of a nation, but the government can also follow changes in it to adjust economic policies, economic sectors, salaries, and social security, thereby ramping up economic development. Machine-learning models play a key role in forecasting and determining the influence of other factors on the US CPI. Consequently, future studies could apply the MARS, SVR, ARDL, and MLR models with these independent variables to predict the CPI of other nations. Furthermore, subsequent studies should consider more significant variables with a larger number of samples, and other machine-learning models should be deployed for the predictive analysis of CPIs in other developed countries to provide an overview of the global economy in terms of improving the accuracy of CPI forecasting. Simultaneously, other appropriate multivariate techniques with accurate forecasting could be enhanced and deployed in future studies to project inflation.

Funding

This research received no external funding.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Tien-Thinh Nguyen: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Hong-**

Giang Nguyen: Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Jen-Yao Lee:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Yu-Lin Wang:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Chien-Shu Tsai:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors appreciate the support from the National Kaohsiung University of Science and Technology, Taiwan.

References

- [1] R.D. Schnepp, J. Richardson, *Consumers and Food Price Inflation*, Congressional Research Service, Library of Congress, 2009.
- [2] S.G. Cecchetti, R. Moreno, D. Mihajek, A. Villar, S.C. Saxena, Monetary Policy and the Measurement of Inflation: Prices, Wages and Expectations, 2010. BIS Paper 49, <https://ssrn.com/abstract=1546992>.
- [3] H. Hassani, A.S. Soofi, A. Zhigljavsky, Predicting inflation dynamics with singular spectrum analysis, *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)* 176 (3) (2013) 743–760, <https://doi.org/10.1111/j.1467-985X.2012.01061.x>.
- [4] R. Edvinsson, J. Söderberg, A consumer price index for Sweden, 1290–2008, *Rev. Income Wealth* 57 (2) (2011) 270–292, <https://doi.org/10.1111/j.1475-4991.2010.00381.x>.
- [5] S. Gumbe, N. Kaseke, Manufacturing firms and hyperinflation-survival options: the case of Zimbabwe manufacturers (2005–2008), *Journal of Management and Marketing Research* 7 (2011) 1.
- [6] R.K. Weaver, *Automatic Government: the Politics of Indexation*, Brookings Institution Press, 2010.
- [7] L. Gao, H. Kim, R. Saba, How do oil price shocks affect consumer prices? *Energy Econ.* 45 (2014) 313–323, <https://doi.org/10.1016/j.eneco.2014.08.001>.
- [8] J.M. Binner, P. Tino, J. Tepper, R. Anderson, B. Jones, G. Kendall, Does money matter in inflation forecasting? *Phys. Stat. Mech. Appl.* 389 (21) (2010) 4793–4808.
- [9] J. Beckmann, R. Czudaj, Oil and gold price dynamics in a multivariate cointegration framework, *Int. Econ. Econ. Pol.* 10 (2013) 453–468, <https://doi.org/10.1007/s10368-013-0237-8>.
- [10] G. Filis, Macro economy, stock market and oil prices: do meaningful relationships exist among their cyclical fluctuations? *Energy Econ.* 32 (4) (2010) 877–886, <https://doi.org/10.1016/j.eneco.2010.03.010>.
- [11] K.W. Chou, Y.H. Tseng, Pass-through of oil prices to CPI inflation in Taiwan, *International Research Journal of Finance and Economics* 69 (69) (2011) 73–83.
- [12] S. Long, J. Liang, Asymmetric and nonlinear pass-through of global crude oil price to China's PPI and CPI inflation, *Economic Research-Ekonomska Istraživanja* 31 (1) (2018) 240–251, <https://doi.org/10.1080/1331677X.2018.1429292>.
- [13] Z. Qianqian, The impact of international oil price fluctuation on China's economy, *Energy Proc.* 5 (2011) 1360–1364, <https://doi.org/10.1016/j.egypro.2011.03.235>.
- [14] M.R. Barakat, S.H. Elgazzar, K.M. Hanafy, Impact of macroeconomic variables on stock markets: evidence from emerging markets, *Int. J. Econ. Finance* 8 (1) (2016) 195–207, <https://doi.org/10.5539/ijef.v8n1p195>. URL.
- [15] B. Bernanke, A. Blinder, The Federal Funds Rate and the Channels of Monetary Transmission, 1990, <https://doi.org/10.3386/w3487>. NBER Working Paper No. 3487.
- [16] I. Mukhlis, I. Hidayah, N.R. Retnasih, Interest rate volatility of the federal funds rate: response of the bank Indonesia and its impact on the Indonesian economic stability, *J. Cent. Bank Theor. Pract.* 9 (1) (2020) 111–133, <https://doi.org/10.2478/jcbtp-2020-0007>.
- [17] M.Q. Khasawneh, The impact of federal fund rate in USA movements on lending rates and consumer price index in Jordan, *Journal of Eastern European and Central Asian Research* 3 (1) (2016) 11, <https://doi.org/10.15549/jeccar.v3i1.117>, 11.
- [18] S.L. Hashim, H. Ramli, N.H. Razali, N.Z. Nordin, Macroeconomic variables affecting the volatility of gold price, *Journal of Global Business and Social Entrepreneurship* 3 (5) (2017) 97–106.
- [19] P. Mali, Fluctuation of gold price in India versus global consumer price index, *Fractals* 22 (01n02) (2014), 1450004, <https://doi.org/10.1142/S0218348X14500042>.
- [20] G. Tkacz, Gold Prices and Inflation, vol. 35, Bank of Canada Working Paper No, 2007, <https://doi.org/10.34989/swp-2007-35>, 2007.
- [21] S. Cheng, Y. Jin, S.P. Harrison, C. Quilodrán-Casas, I.C. Prentice, Y.K. Guo, R. Arcucci, Parameter flexible wildfire prediction using machine learning techniques: forward and inverse modelling, *Rem. Sens.* 14 (13) (2022) 3228, <https://doi.org/10.3390/rs14133228>.
- [22] J. Ding, C. Zhang, D. Li, A.K. Sangaiah, Hyperautomation for Air Quality Evaluations: A Perspective of Evidential Three-Way Decision-Making, *Cognitive Computation*, 2023, pp. 1–17, <https://doi.org/10.1007/s12559-022-10101-8>.
- [23] H. Gong, S. Cheng, Z. Chen, Q. Li, C. Quilodrán-Casas, D. Xiao, R. Arcucci, An efficient digital twin based on machine learning SVD autoencoder and generalised latent assimilation for nuclear reactor physics, *Ann. Nucl. Energy* 179 (2022), 109431, <https://doi.org/10.1016/j.anucene.2022.109431>.
- [24] L. Schoppa, M. Disse, S. Bachmair, Evaluating the performance of random forest for large-scale flood discharge simulation, *J. Hydrol.* 590 (2020), 125531, <https://doi.org/10.1016/j.jhydrol.2020.125531>.
- [25] C. Zhu, X. Ma, C. Zhang, W. Ding, J. Zhan, Information granules-based long-term forecasting of time series via BPNN under three-way decision framework, *Inf. Sci.* 634 (2023) 696–715, <https://doi.org/10.1016/j.ins.2023.03.133>.
- [26] Z. Affes, R. Bentati-Kaffel, Forecast bankruptcy using a blend of clustering and MARS model: case of US banks, *Ann. Oper. Res.* 281 (1) (2019) 27–64, <https://doi.org/10.1007/s10479-018-2845-8>.
- [27] T.S. Lee, C.C. Chiu, Y.C. Chou, C.J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, *Comput. Stat. Data Anal.* 50 (4) (2006) 1113–1130, <https://doi.org/10.1016/j.csda.2004.11.006>.
- [28] S.S. Roy, R. Roy, V.E. Balas, Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM, *Renew. Sustain. Energy Rev.* 82 (2018) 4256–4268, <https://doi.org/10.1016/j.rser.2017.05.249>.
- [29] G. Nalcaci, A. Özmen, G.W. Weber, Long-term load forecasting: models based on MARS, ANN and LR methods, *Cent. Eur. J. Oper. Res.* 27 (4) (2019) 1033–1049, <https://doi.org/10.1007/s10100-018-0531-1>.
- [30] M.Y. Cheng, M.T. Cao, Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines, *Appl. Soft Comput.* 22 (2014) 178–188, <https://doi.org/10.1016/j.asoc.2014.05.015>.
- [31] C.B. Santiago, J.Y. Guo, M.S. Sigman, Predictive and mechanistic multivariate linear regression models for reaction development, *Chem. Sci.* 9 (9) (2018) 2398–2412, <https://doi.org/10.1039/C7SC04679K>.
- [32] G. Forkuor, O.K. Houkpatin, G. Welp, M. Thiel, High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models, *PLoS One* 12 (1) (2017), e0170478, <https://doi.org/10.1371/journal.pone.0170478>.

- [33] J. Riofrío, O. Chang, E.J. Revelo-Fuelagán, D.H. Peluffo-Ordóñez, Forecasting the Consumer Price Index (CPI) of Ecuador: a comparative study of predictive models, *Int. J. Adv. Sci. Eng. Inf. Technol.* 10 (3) (2020) 1078–1084.
- [34] E. Gjika, L. Puka, O. Zaçaj, Forecasting consumer price index (CPI) using time series models and multi regression models (Albania case study), in: “10th International Scientific Conference” Business and Management, 2018, pp. 3–4, <https://doi.org/10.3846/bm.2018.51>.
- [35] C. Lidima, Modelling and forecasting inflation rate in Kenya using SARIMA and holt-winters triple exponential smoothing, *Am. J. Theor. Appl. Stat.* 6 (3) (2017) 161–169, <https://doi.org/10.11648/j.ajtas.20170603.15>.
- [36] V. Ülke, A. Sahin, A. Subasi, A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the USA, *Neural Comput. Appl.* 30 (5) (2018) 1519–1527, <https://doi.org/10.1007/s00521-016-2766-x>.
- [37] R. Li, X. Han, Q. Wang, Do technical differences lead to a widening gap in China's regional carbon emissions efficiency? Evidence from a combination of LMDI and PDA approach, *Renew. Sustain. Energy Rev.* 182 (2023), 113361, <https://doi.org/10.1016/j.rser.2023.113361>.
- [38] R. Li, L. Li, Q. Wang, The impact of energy efficiency on carbon emissions: evidence from the transportation sector in Chinese 30 provinces, *Sustain. Cities Soc.* 82 (2022), 103880, <https://doi.org/10.1016/j.scs.2022.103880>.
- [39] Q. Wang, J. Sun, R. Li, U.K. Pata, Linking trade openness to load capacity factor: the threshold effects of natural resource rent and corruption control, *Gondwana Res.* (2023), <https://doi.org/10.1016/j.gr.2023.05.016>.
- [40] Q. Wang, J. Sun, U.K. Pata, R. Li, M.T. Kartal, Digital economy and carbon dioxide emissions: examining the role of threshold variables, *Geosci. Front.* (2023), 101644, <https://doi.org/10.1016/j.gsf.2023.101644>.
- [41] Q. Wang, L. Wang, R. Li, Trade Openness Helps Move towards Carbon Neutrality Insight from 114 Countries, *Sustainable Development*, 2023, <https://doi.org/10.1002/sd.2720>.
- [42] Q. Wang, X. Wang, R. Li, Does urbanization redefine the environmental Kuznets curve? An empirical analysis of 134 Countries, *Sustain. Cities Soc.* 76 (2022), 103382, <https://doi.org/10.1016/j.scs.2021.103382>.
- [43] Q. Wang, F. Zhang, R. Li, Free Trade and Carbon Emissions Revisited: the Asymmetric Impacts of Trade Diversification and Trade Openness, *Sustainable Development*, 2023, <https://doi.org/10.1002/sd.2703>.
- [44] B. Kalayci, A. Özmen, G.W. Weber, Mutual relevance of investor sentiment and finance by modeling coupled stochastic systems with MARS, *Ann. Oper. Res.* 295 (1) (2020) 183–206, <https://doi.org/10.1007/s10479-020-03757-8>.
- [45] M. Yıldız, L. Özdemir, Determination of the sensitivity of stock index to macroeconomic and psychological factors by MARS method, in: Insurance and Risk Management for Disruptions in Social, Economic and Environmental Systems: Decision and Control Allocations within New Domains of Risk, Emerald Publishing Limited, 2022, <https://doi.org/10.1108/978-1-80117-139-720211005>.
- [46] L.Y. Chang, Analysis of bilateral air passenger flows: a non-parametric multivariate adaptive regression spline approach, *J. Air Transport. Manag.* 34 (2014) 123–130, <https://doi.org/10.1016/j.jairtraman.2013.09.003>.
- [47] D. Cogoljević, M. Gavrilović, M. Roganović, I. Matić, I. Piljan, Analyzing of consumer price index influence on inflation by multiple linear regression, *Phys. Stat. Mech. Appl.* 505 (2018) 941–944, <https://doi.org/10.1016/j.physa.2018.04.014>.
- [48] Y. Wang, B. Wang, X. Zhang, A new application of the support vector regression on the construction of financial conditions index to CPI prediction, *Proc. Comput. Sci.* 9 (2012) 1263–1272, <https://doi.org/10.1016/j.procs.2012.04.138>.
- [49] R. Peirano, W. Kristjanpoller, M.C. Minutolo, Forecasting inflation in Latin American countries using a SARIMA-LSTM combination, *Soft Comput.* 25 (16) (2021) 10851–10862, <https://doi.org/10.1007/s00500-021-06016-5>.
- [50] M. Simionescu, Econometrics of sentiments-sentometrics and machine learning: the improvement of inflation predictions in Romania using sentiment analysis, *Technol. Forecast. Soc. Change* 182 (2022), 121867, <https://doi.org/10.1016/j.techfore.2022.121867>.
- [51] R. Maldeni, M.A. Mascrenhe, A machine learning approach to CCPI-based inflation prediction, in: Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, vol. 2, Springer Singapore, Singapore, 2021, pp. 567–575, https://doi.org/10.1007/978-981-16-2380-6_50.
- [52] O. Barkan, J. Benchimol, I. Caspi, E. Cohen, A. Hammer, N. Koenigstein, Forecasting CPI inflation components with hierarchical recurrent neural networks, *Int. J. Forecast.* (2022), <https://doi.org/10.1016/j.ijforecast.2022.04.009>.
- [53] A.A. Salisu, K.O. Isah, Predicting US inflation: evidence from a new approach, *Econ. Modell.* 71 (2018) 134–158, <https://doi.org/10.1016/j.economod.2017.12.008>.
- [54] A. Özmen, Y. Yılmaz, G.W. Weber, Natural gas consumption forecast with MARS and CMARS models for residential users, *Energy Econ.* 70 (2018) 357–381, <https://doi.org/10.1016/j.eneco.2018.01.022>.
- [55] R.A. Groeneveld, G. Meeden, Measuring skewness and kurtosis, *J. Roy. Stat. Soc.: Series D (The Statistician)* 33 (4) (1984) 391–399, <https://doi.org/10.2307/2987742>.
- [56] D.A. Dickey, W.A. Fuller, Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica: J. Econom. Soc.* (1981) 1057–1072, <https://doi.org/10.2307/1912517>.
- [57] P.C. Phillips, P. Perron, Testing for a unit root in time series regression, *Biometrika* 75 (2) (1988) 335–346, <https://doi.org/10.1093/biomet/75.2.335>.
- [58] M.H. Pesaran, A simple panel unit root test in the presence of cross-section dependence, *J. Appl. Econom.* 22 (2) (2007) 265–312, <https://doi.org/10.1002/jae.951>.
- [59] A. Levin, C.F. Lin, C.S.J. Chu, Unit root tests in panel data: asymptotic and finite-sample properties, *J. Econom.* 108 (1) (2002) 1–24, [https://doi.org/10.1016/S0304-4076\(01\)00098-7](https://doi.org/10.1016/S0304-4076(01)00098-7).
- [60] C. Toraman, C. Başarır, The long run relationship between stock market capitalization rate and interest rate: Co-integration approach, *Procedia-Social and Behavioral Sciences* 143 (2014) 1070–1073, <https://doi.org/10.1016/j.sbspro.2014.07.557>.
- [61] F. Bashir, S. Nawaz, K. Yasin, U. Khursheed, J. Khan, M.J. Qureshi, Determinants of inflation in Pakistan: an econometric analysis using Johansen co-integration approach, *Aust. J. Bus. Manag. Res.* 1 (5) (2011) 71–82, <https://mpra.ub.uni-muenchen.de/id/eprint/106870>.
- [62] Z. Wang, R.S. Srinivasan, A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models, *Renew. Sustain. Energy Rev.* 75 (2017) 796–808, <https://doi.org/10.1016/j.rser.2016.10.079>.
- [63] M. Tranmer, M. Elliot, Multiple linear regression, *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5 (5) (2008) 1–5.
- [64] D. Maulud, A.M. Abdulazeez, A review on linear regression comprehensive in machine learning, *Journal of Applied Science and Technology Trends* 1 (4) (2020) 140–147, <https://doi.org/10.38094/jastt1457>.
- [65] D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, 2010.
- [66] S. Sekulic, B.R. Kowalski, MARS: a tutorial, *J. Chemometr.* 6 (4) (1992) 199–216, <https://doi.org/10.1002/cem.1180060405>.
- [67] J. Fan, L. Wu, X. Ma, H. Zhou, F. Zhang, Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions, *Renew. Energy* 145 (2020) 2034–2045, <https://doi.org/10.1016/j.renene.2019.07.104>.
- [68] V.N. Sharda, S.O. Prasher, R.M. Patel, P.R. Ojasvi, C. Prakash, Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data/Performances de régressions par splines multiples et adaptives (MARS) pour la prévision d'écoulement au sein de micro-bassins versants Himalayens d'altitudes intermédiaires avec peu de données, *Hydrol. Sci. J.* 53 (6) (2008) 1165–1175, <https://doi.org/10.1623/hysj.53.6.1165>.
- [69] A. Parsaei, A.H. Haghiahi, Mathematical expression of discharge capacity of compound open channels using MARS technique, *J. Earth Syst. Sci.* 126 (2017) 1–15, <https://doi.org/10.1007/s12040-017-0807-1>.
- [70] N.O. Attoh-Okine, K. Cooger, S. Mensah, Multivariate adaptive regression (MARS) and hinged hyperplanes (HHP) for doweled pavement performance modeling, *Construct. Build. Mater.* 23 (9) (2009) 3020–3023, <https://doi.org/10.1016/j.conbuildmat.2009.04.010>.
- [71] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Software* 33 (1) (2010) 1.
- [72] M. LeBlanc, R. Tibshirani, Adaptive principal surfaces, *J. Am. Stat. Assoc.* 89 (425) (1994) 53–64, <https://doi.org/10.1080/01621459.1994.10476445>.

- [73] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.* 14 (3) (2004) 199–222, <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
- [74] B. Keshtegar, M.L. Nehdi, N.T. Trung, R. Kolahchi, Predicting load capacity of shear walls using SVR-RSM model, *Appl. Soft Comput.* 112 (2021), 107739, <https://doi.org/10.1016/j.asoc.2021.107739>.
- [75] H. Zhang, X. Cheng, Y. Li, X. Du, Prediction of failure modes, strength, and deformation capacity of RC shear walls through machine learning, *J. Build. Eng.* 50 (2022), 104145, <https://doi.org/10.1016/j.jobe.2022.104145>.
- [76] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: a review, *Renew. Sustain. Energy Rev.* 23 (2013) 272–288, <https://doi.org/10.1016/j.rser.2013.03.004>.
- [77] Z. Xu, Y. Gao, Y. Jin, Application of an optimized SVR model of machine learning, *International Journal of Multimedia and Ubiquitous Engineering* 9 (6) (2014) 67–80, <https://doi.org/10.14257/ijmuae.2014.9.6.08>.
- [78] M. Awad, R. Khanna, Support vector machines for classification, in: *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Apress, Berkeley, CA, 2015, pp. 39–66.
- [79] N.M. Odhiambo, Energy consumption and economic growth nexus in Tanzania: an ARDL bounds testing approach, *Energy Pol.* 37 (2) (2009) 617–622, <https://doi.org/10.1016/j.enpol.2008.09.077>.
- [80] M.H. Pesaran, Y. Shin, *An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis*. Cambridge, Department of Applied Economics, University of Cambridge, UK, 1995, p. 9514.
- [81] M.H. Pesaran, Y. Shin, R.J. Smith, Bounds testing approaches to the analysis of level relationships, *J. Appl. Econom.* 16 (3) (2001) 289–326, <https://doi.org/10.1002/jae.616>.
- [82] Q. Kong, D. Peng, Y. Ni, X. Jiang, Z. Wang, Trade openness and economic growth quality of China: empirical analysis using ARDL model, *Finance Res. Lett.* 38 (2021), 101488, <https://doi.org/10.1016/j.frl.2020.101488>.
- [83] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petković, C. Sudheer, A support vector machine–firefly algorithm-based model for global solar radiation prediction, *Sol. Energy* 115 (2015) 632–644, <https://doi.org/10.1016/j.solener.2015.03.015>.
- [84] J.Y. Lee, T.T. Nguyen, H.G. Nguyen, J.Y. Lee, Towards predictive crude oil purchase: a case study in the USA and europe, *Energies* 15 (11) (2022) 4003, <https://doi.org/10.3390/en15114003>.
- [85] M.S. Bakay, Ü. Ağbulut, Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms, *J. Clean. Prod.* 285 (2021), 125324, <https://doi.org/10.1016/j.jclepro.2020.125324>.
- [86] A.E. Gürel, Ü. Ağbulut, Y. Biçen, Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation, *J. Clean. Prod.* 277 (2020), 122353, <https://doi.org/10.1016/j.jclepro.2020.122353>.
- [87] A.O. Ali, J. Mohamed, The optimal forecast model for consumer price index of Puntland State, Somalia, *Qual. Quantity* 56 (6) (2022) 4549–4572, <https://doi.org/10.1007/s11135-022-01328-6>.
- [88] A. Özbayrak, M.K. Ali, H. Çatakoglu, Buckling load estimation using multiple linear regression analysis and multigene genetic programming method in cantilever beams with transverse stiffeners, *Arabian J. Sci. Eng.* 48 (4) (2023) 5347–5370, <https://doi.org/10.1007/s13369-022-07445-6>.
- [89] Y. Zouzou, H. Citakoglu, General and regional cross-station assessment of machine learning models for estimating reference evapotranspiration, *Acta Geophys.* 71 (2) (2023) 927–947, <https://doi.org/10.1007/s11600-022-00939-9>.
- [90] K.E. Taylor, Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res. Atmos.* 106 (D7) (2001) 7183–7192, <https://doi.org/10.1029/2000JD900719>.
- [91] V. Demir, Enhancing monthly lake levels forecasting using heuristic regression techniques with periodicity data component: application of Lake Michigan, *Theor. Appl. Climatol.* 148 (3–4) (2022) 915–929, <https://doi.org/10.1007/s00704-022-03982-0>.
- [92] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [93] M.A. Sahraei, H. Duman, M.Y. Çodur, E. Eyduran, Prediction of transportation energy demand: multivariate adaptive regression splines, *Energy* 224 (2021), 120090, <https://doi.org/10.1016/j.energy.2021.120090>.
- [94] C. Haouraji, B. Mounir, I. Mounir, A. Farchi, A correlative approach, combining energy consumption, urbanization and GDP, for modeling and forecasting Morocco's residential energy consumption, *International Journal of Energy and Environmental Engineering* 11 (1) (2020) 163–176, <https://doi.org/10.1007/s40095-020-00336-2>.
- [95] D. Enke, N. Mehdiyev, A hybrid neuro-fuzzy model to forecast inflation, *Proc. Comput. Sci.* 36 (2014) 254–260, <https://doi.org/10.1016/j.procs.2014.09.088>.
- [96] C. Yang, S. Guo, Inflation Prediction Method Based on Deep Learning, *Computational Intelligence and Neuroscience*, 2021, <https://doi.org/10.1155/2021/1071145>, 2021.
- [97] J. Gama, I. Ziobaité, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, *ACM Comput. Surv.* 46 (4) (2014) 1–37, <https://doi.org/10.1145/2523813>.