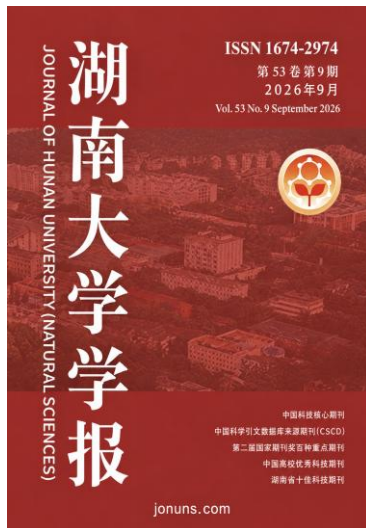




Journal of Hunan University (Natural Sciences)

Vol. 53 No. 9
September 2026

Available online at
<https://jonuns.com>



ELSEVIER
Scopus



Clarivate
WEB OF SCIENCE

Open Access Article

 <https://doi.org/10.55463/issn.1674-2974.53.9.1>

Machine Learning versus Deep Learning in Macroeconomic Forecasting: A Global GDP Growth Application

Fazale Naeem^{1*}, Muhammad Naveed¹, Muhammad Fayez¹, Santosh Prasad Sah¹

¹School of Economics and Trade, Hunan University, Changsha, China

* Corresponding author: fnaemkhan@hnu.edu.cn, darghainaeem@gmail.com

Article History:

Received: August 19, 2026

Revised: September 25, 2026

Accepted: October 8, 2026

Published: October 20, 2026

Abstract: This study compares regularized linear, tree-based ensemble, boosting, and feed-forward neural-network approaches for forecasting country-level annual GDP growth using a global panel covering 213 countries from 2010 to 2023. A chronological out-of-sample evaluation is employed, with models trained on observations through 2021 and tested on 401 country-year observations from 206 countries during 2022–2023. Ridge regression, Random Forest, Histogram-Based Gradient Boosting, AdaBoost, and a multi-layer perceptron (MLP) are evaluated using root mean squared error (RMSE). Ridge regression records the lowest observed test RMSE (5.50), followed by Histogram-Based Gradient Boosting (6.28), Random Forest (6.75), AdaBoost (7.13), and the MLP (8.12). AdaBoost is treated as a classical boosting-based machine-learning method, whereas the MLP serves as the neural-network benchmark. The baseline models use one-year lags, while AdaBoost and the MLP use one-to-three-year lag windows. Consequently, the observed ranking reflects differences in both model specifications and information sets and should not be interpreted as a strictly like-for-like comparison or evidence of statistical superiority. Within this annual panel, predictor set, and 2022–2023 evaluation period, greater model complexity is not associated with lower forecast errors. The findings highlight the practical importance of regularization, model parsimony, and chronological evaluation in country-level macroeconomic forecasting.



Copyright: © 2026 by the authors. Licensee JHU

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License
<http://creativecommons.org/licenses/by/4.0/>

Keywords: GDP growth forecasting, Machine learning, Deep learning, Macroeconomic forecasting, Global panel data, Out-of-sample evaluation, Model comparison.

机器学习与深度学习在宏观经济预测中的比较研究：基于全球 GDP 增长的实证分析

摘要：准确预测国内生产总值（GDP）增长率对于宏观经济分析、政策制定以及经济风险评估具有重要意义。本文基于覆盖213个国家、时间跨度为2010_2023年的全球宏观经济面板数据，系统比较了机器学习与深度学习方法在年度GDP增长率预测中的表现。研究采用严格的时间序列外样本评估框架，以2010_2021年数据作为训练样本，并以2022_2023年数据作为测试样本，从而模拟真实预测环境并避免前视偏差。在模型设定方面，本文选取正则化线性模型、基于树的机器学习模型以及深度学习风格模型进行比较分析，并以均方根误差（RMSE）作为主要预测精度评价指标。实证结果表明，岭回归模型在外样本预测中取得最低RMSE，其预测性能优于随机森林、梯度提升、AdaBoost以及多层感知机等更复杂的非线性模型。研究结果显示，在全球年度宏观经济预测情境下，模型复杂度的提高并不必然带来预测精度的改善。相反，正则化处理、模型简约性、合理的动态特征构造以及严格的时间序列评估设计，对于提升宏观经济预测的稳定性和可靠性更为关键。本文为机器学习与深度学习方法在宏观经济预测领域的应用提供了经验证据，并对在低频、高噪声且存在结构性冲击的宏观经济数据环境中盲目采用复杂深度学习模型提出了谨慎启示。

关键词：GDP 增长预测,机器学习,深度学习,宏观经济预测,全球面板数据,外样本评估

1. Introduction

Accurate forecasting of gross domestic product (GDP) growth is important for macroeconomic analysis and policy formulation. Governments, central banks, and international institutions use GDP growth forecasts to support fiscal planning, monetary policy decisions, and development strategies. Large forecast errors may contribute to inappropriate policy responses, resource misallocation, and heightened uncertainty, particularly during periods of economic disruption. Improving the reliability of GDP growth forecasts therefore remains an important objective in applied macroeconomics.

Traditional macroeconomic forecasting has relied on econometric methods such as autoregressive models, Vector Autoregressions (VARs), and factor models. These approaches provide interpretable frameworks for representing economic dynamics, but they may face difficulties when the predictor set is highly dimensional, relationships change over time, or the data contain complex interactions and structural shocks. The increasing availability of cross-country macroeconomic indicators has consequently encouraged researchers to

examine alternative forecasting methods capable of using broader information sets.

Machine-learning methods have gained prominence in economic forecasting because they offer tools for regularization, flexible function approximation, and the detection of interactions among predictors. Regularized regression can stabilize estimates when predictors are strongly correlated, while tree-based ensembles can represent nonlinearities and threshold effects without requiring them to be specified in advance. Nevertheless, empirical evidence remains mixed: the relative performance of different methods depends on the target variable, data frequency, sample size, predictor set, tuning procedure, and evaluation period.

Neural networks have achieved considerable success in fields such as computer vision, natural-language processing, and high-frequency financial analysis. These developments have created interest in their application to macroeconomic forecasting. However, annual macroeconomic panels differ substantially from the large and high-frequency datasets in which neural networks often perform well. Country-

level annual series are comparatively short, noisy, heterogeneous, and affected by infrequent but substantial shocks. These characteristics may limit the forecasting benefits of highly parameterized neural-network models.

Several gaps motivate the present study. First, many macroeconomic forecasting studies concentrate on individual countries or limited regional samples, which restricts the assessment of model performance across heterogeneous economies. Second, some evaluations employ random train-test splits or emphasize in-sample fit, even though such designs may provide overly optimistic estimates of forecasting performance. Third, comparisons among regularized linear models, tree-based ensembles, boosting methods, and neural networks remain relatively limited in broad country-level annual panels.

This study addresses these gaps by comparing several predictive specifications for forecasting country-level annual GDP growth using a panel of 213 countries covering 2010–2023. Observations through 2021 form the training period, while observations from 2022–2023 are reserved for out-of-sample evaluation. This chronological design more closely reflects a forecasting application than a random split and is intended to reduce the risk of look-ahead bias.

The empirical analysis distinguishes four methodological groups. Ridge regression represents the regularized linear approach. Random Forest and Histogram-Based Gradient Boosting represent tree-based machine-learning methods. AdaBoost is treated separately as a classical boosting-based ensemble method. A feed-forward multi-layer perceptron (MLP) serves as the neural-network benchmark. Root mean squared error (RMSE) is used as the primary measure of test-period forecasting accuracy.

The implemented model specifications do not use identical lagged information sets. Ridge regression, Random Forest, and Histogram-Based Gradient Boosting use one-year lags, whereas AdaBoost and the MLP use lag windows of one to three years. Consequently, comparisons between these groups reflect differences in both the algorithms and their input information sets. The results should therefore be interpreted as a descriptive comparison of the implemented forecasting specifications rather than a strictly like-for-like test of algorithmic performance.

Ridge regression records the lowest observed test RMSE of 5.50, followed by Histogram-Based Gradient Boosting at 6.28, Random Forest at 6.75, AdaBoost at 7.13, and the MLP at 8.12. These rankings indicate that greater model flexibility is not associated with lower forecast errors in the particular annual panel and evaluation period considered. However, because the test period contains only two annual forecast dates and the specifications use different lag structures, the findings do not demonstrate statistical superiority or establish

that regularized linear models generally dominate tree-based or neural-network methods.

The study contributes to the literature by providing a broad country-level comparison under a chronological evaluation design and by illustrating the practical importance of model parsimony and regularization in noisy, low-frequency macroeconomic data. Its conclusions are limited to the selected predictors, preprocessing choices, lag specifications, country coverage, and 2022–2023 test period. The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the data and methodology. Section 4 presents the empirical results, Section 5 discusses their interpretation and limitations, and Section 6 concludes.

2. Literature Review

Macroeconomic forecasting has increasingly adopted data-rich methods to extract useful signals from large sets of economic indicators. Diffusion-index models demonstrate that large macroeconomic panels can be summarized through a small number of latent factors, improving forecasts of real activity and inflation relative to models based on limited information sets [1]. Nowcasting research further shows that forecast accuracy can benefit from incorporating newly released information as it becomes available [2]. Because the usefulness of factor-based models depends on the number of retained components, formal information criteria have been developed to balance information extraction against the risk of introducing additional estimation noise [3]. These studies establish the importance of broad predictor sets, dimensionality control, and realistic out-of-sample evaluation.

Shrinkage provides an alternative to factor extraction when many correlated predictors are available. Bayesian shrinkage can stabilize high-dimensional forecasting regressions while retaining predictors in their original variable space [4]. Similar principles underlie large Bayesian Vector Autoregressions (BVARs), which control parameter proliferation through structured priors [5]. The performance of these models is sensitive to lag length, prior tightness, and variable selection, making transparent specification and robustness analysis important [6]. Dynamic model averaging extends this framework by allowing model probabilities and predictor relevance to change over time [7]. Mixed-frequency VAR models address another practical issue by combining variables observed at different frequencies and accounting for the timing of information releases [8]. Although these approaches are not estimated in the present study, they illustrate the importance of regularization, temporal structure, and carefully defined information sets.

Reliable model comparison requires more than reporting the smallest forecast error. Tests of equal

predictive accuracy provide a formal basis for assessing whether differences between forecast sequences are statistically meaningful [9]. Related asymptotic results clarify the conditions under which observed differences in predictive ability may be distinguished from sampling variation [10]. These methods generally require sufficiently long sequences of forecast errors. Consequently, their applicability is limited when an evaluation period contains only a small number of annual forecast dates, as in the present study.

The growth of large economic datasets has encouraged the use of machine-learning methods in empirical economics [11,12]. Machine learning is particularly relevant to prediction problems because it emphasizes regularization, cross-validation, and flexible function approximation [13]. High-dimensional econometric methods also demonstrate how disciplined variable selection and regularization can reduce instability when the number of potential predictors is large [14]. Surveys of machine learning in economics emphasize both its strengths, including the ability to represent interactions and nonlinearities, and its risks, including overfitting, instability under structural change, and limited interpretability [15]. Ridge regression is especially relevant when predictors are strongly correlated because it stabilizes coefficient estimates by shrinking them toward zero [16]. Related penalized methods, such as the lasso, combine shrinkage with variable selection [17]. These findings support the use of regularized linear models as serious forecasting benchmarks rather than merely simple alternatives to more complex algorithms.

Tree-based ensembles provide flexible methods for representing nonlinearities, interactions, and threshold effects. Random Forest combines predictions from multiple decorrelated decision trees, reducing the variance associated with a single tree [18]. Its performance depends on choices such as tree depth, feature selection, sampling, and the number of trees [19]. Gradient boosting constructs an additive predictive model by sequentially fitting weak learners to previous errors [20]. AdaBoost similarly combines weak learners through sequential reweighting of observations [21]. Importantly, AdaBoost is a classical ensemble machine-learning method and should not be classified as a deep-learning model. Although tree-based and boosting methods can approximate complex relationships, they require careful tuning and time-respecting validation because their flexibility does not guarantee improved out-of-sample performance.

Forecast performance may also depend on the forecast horizon and the composition of the predictor set. Direct and iterated multi-step forecasting methods have different advantages, and neither approach dominates in all settings [23]. Financial indicators can provide forward-looking information about economic activity; for example, the yield curve has demonstrated

predictive content for future GDP growth [22]. These findings indicate that differences in forecasting accuracy can arise not only from the selected algorithm but also from the information set, lag structure, forecast horizon, and predictor transformations. Comparisons across algorithms are therefore most informative when these design elements are harmonized.

Forecast-accuracy measures summarize different aspects of predictive performance. RMSE places relatively greater weight on large errors, while mean absolute error is less sensitive to extreme observations [24]. The appropriate measure depends on the forecasting objective and the distribution of the target variable. Evidence from large forecasting competitions also shows that simple statistical models and forecast combinations can be difficult to improve upon consistently [25]. These results underline the importance of including simple benchmarks and avoiding the assumption that greater algorithmic complexity necessarily produces more accurate forecasts.

Neural networks have a long history in forecasting applications. Early studies show that feed-forward networks can represent nonlinear relationships but require careful architecture selection, data scaling, optimization, and regularization [26]. Universal-approximation results provide a theoretical basis for their flexibility, although they do not imply superior empirical forecasting performance [27]. Modern deep learning extends these ideas through deeper architectures, improved optimization, and representation learning [28]. Recurrent models such as long short-term memory (LSTM) networks are designed to represent longer temporal dependencies [29]. However, their advantages generally depend on the availability of sufficiently long and informative time series. The present study does not estimate an LSTM or another recurrent architecture; its neural-network evidence is limited to a feed-forward MLP. The results should therefore not be interpreted as a comprehensive evaluation of deep learning.

Recent empirical research demonstrates that machine learning can improve some macroeconomic forecasts, particularly when large predictor sets contain useful nonlinear information. Machine-learning methods have produced gains in data-rich inflation-forecasting applications [30]. Broader evaluations nevertheless show that improvements vary across economic variables, forecast horizons, model specifications, and evaluation periods [31]. The evidence therefore supports conditional rather than universal claims about the benefits of machine learning. Complex methods should be evaluated against strong regularized and naive benchmarks using chronological validation procedures.

The predictive advantage of a model may also vary across economic conditions. Tests of conditional predictive ability allow relative forecast performance to

differ between states, such as expansions and recessions [32]. Specialized procedures are also required when a simpler model is nested within a more flexible alternative [33]. Data resources such as FRED-MD have facilitated systematic forecasting comparisons for individual economies using large monthly predictor sets [34]. Cross-country forecasting introduces additional challenges because economies are connected through trade, finance, and common shocks. Global macroeconomic models explicitly represent these interdependencies [35]. The present study instead forecasts country-level annual GDP growth using a global panel and does not explicitly model international spillovers, which should be recognized when interpreting its results.

The literature suggests three conclusions relevant to this study. First, regularization can be highly valuable when macroeconomic predictors are correlated and the available time dimension is limited. Second, tree-based ensembles, boosting methods, and neural networks offer greater functional flexibility, but their forecasting performance depends on the available data, tuning procedure, lag design, and evaluation period. Third, credible comparisons require chronological testing, common information sets where feasible, appropriate benchmarks, and cautious interpretation of observed error differences. Building on these insights, the present study compares Ridge regression, Random Forest, Histogram-Based Gradient Boosting, AdaBoost, and an MLP using country-level annual GDP growth data. Because the implemented model groups use different lag structures, their results are interpreted as a descriptive comparison of forecasting specifications rather than a strictly like-for-like test of algorithms.

3. Methodology

3.1 Data Description and Sample Construction

This study uses an unbalanced global macroeconomic panel covering 213 countries from 2010 to 2023. The unit of observation is the country–year, which allows the analysis to incorporate both cross-country differences and changes in macroeconomic conditions over time. The complete dataset contains 2,912 country–year observations. The panel is unbalanced because the availability of macroeconomic indicators varies across countries and years.

The dependent variable is annual real GDP growth, measured as a percentage. Accordingly, the study forecasts annual GDP growth at the country level rather than a single aggregate global GDP growth rate. Annual data are used because several fiscal, structural, and external-sector indicators are not consistently available at higher frequencies for a broad sample of countries.

The dataset was assembled from World Bank Open Data/World Development Indicators and a

secondary Kaggle dataset described as containing World Bank indicators. However, the original data-extraction files and processing code are no longer available. Consequently, the exact extraction date, the allocation of individual variables between the two sources, the World Bank indicator codes, and any processing performed by the Kaggle provider cannot be verified retrospectively. This limitation should be considered when assessing the reproducibility of the study. No additional provenance details have been introduced where they could not be confirmed from the retained research materials.

The sample is divided chronologically to preserve the forecasting structure. Observations from 2010 to 2021 constitute the training sample, while observations from 2022 and 2023 constitute the out-of-sample test period. The training sample contains 2,511 observations from 213 countries, whereas the test sample contains 401 observations from 206 countries. Seven countries represented in the full dataset do not appear in the final test sample because complete usable test-period observations were unavailable after construction of the modeling variables.

This chronological division is preferable to a random train–test split because it ensures that earlier observations are used to predict later outcomes. It therefore more closely reflects a real forecasting exercise and reduces the risk of look-ahead bias.

3.2 Explanatory Variables and Economic Rationale

The explanatory variables were selected according to macroeconomic theory, previous empirical growth research, policy relevance, and availability across a large number of countries. The predictor set includes the following variables:

- **Inflation, measured by the consumer price index (%):** represents price stability and nominal economic conditions that may affect consumption, investment, and purchasing power.
- **Unemployment rate (%):** captures labor-market slack and cyclical economic conditions.
- **Real interest rate (%):** represents the real cost of borrowing and may influence consumption, investment, and capital formation.
- **Public debt (% of GDP):** reflects fiscal obligations and potential sovereign financing constraints.
- **Government expenditure (% of GDP):** captures the size and demand-side stance of fiscal policy.
- **Government revenue (% of GDP):** represents government fiscal capacity and the operation of automatic stabilizers.
- **Tax revenue (% of GDP):** reflects the scale and structure of domestic fiscal financing.

- **Current account balance (% of GDP):** captures external-sector conditions, international financing requirements, and potential external imbalances.

- **GDP per capita (current USD):** serves as an indicator of a country's income and development level.

GDP per capita is included in nominal current US dollars and was not logarithmically transformed in the retained analysis. Because this variable is strongly right-skewed and affected by differences in country size, price levels, and exchange rates, its use in untransformed form is a limitation. A logarithmic transformation could reduce the influence of extreme observations and should be considered in future robustness analyses.

The selected indicators provide a broad and policy-relevant information set, but they do not capture every determinant of GDP growth. Institutional quality, political instability, natural disasters, geopolitical events, and other country-specific shocks are not directly represented. The results should therefore be interpreted as conditional on the predictor set available in this study.

3.3 Exploratory Data Analysis

Exploratory data analysis was conducted before model estimation to examine the structure, quality, and distributional characteristics of the panel. This analysis included descriptive statistics, the distribution of annual GDP growth, missing-value patterns, cross-country and time variation, and correlations among the explanatory variables.

The descriptive analysis indicates substantial heterogeneity across countries and years. GDP growth contains several extreme observations associated with recessions, recoveries, and major economic disruptions. The cross-country average of annual GDP growth also changes considerably over the study period, particularly around the COVID-19 shock. This cross-country average is used only as a descriptive summary and should not be interpreted as an expenditure-weighted measure of aggregate global GDP growth.

Correlation analysis was used to identify linear associations and possible multicollinearity among the predictors. Weak contemporaneous correlations between individual predictors and GDP growth do not establish the existence of nonlinear relationships. They indicate only that simple pairwise linear associations are limited. The inclusion of nonlinear models is therefore treated as an empirical comparison rather than as a conclusion derived from the correlation analysis.

The reported distribution of public debt should be interpreted carefully. In the descriptive table, the 25th percentile, median, and 75th percentile are each reported as 49.16. Because the original data and code are unavailable, these values could not be independently

recalculated during revision. The repeated value may reflect extensive median imputation or a reporting problem and is acknowledged as an unresolved data-verification limitation.

3.4 Dynamic Feature Engineering and Lag Specification

Macroeconomic forecasting requires predictors that would have been available before the outcome being forecast. The study therefore uses lagged values of GDP growth and the explanatory variables. Lagging the predictors helps represent persistence and delayed macroeconomic effects while reducing reliance on contemporaneous information that may not have been available at the time of forecasting.

Lagged variables were constructed separately within each country so that information from one country was not incorrectly transferred to another. The lagged GDP-growth variable captures persistence in economic activity, while lagged fiscal, monetary, labor-market, and external-sector indicators allow the models to represent delayed relationships between macroeconomic conditions and subsequent growth.

The compared models did not use an identical lagged information set. Ridge regression, Random Forest, and Histogram-Based Gradient Boosting used one-year lags. AdaBoost and the Multi-Layer Perceptron used lag windows covering one to three years. The wider lag window was originally intended to provide the more flexible models with additional temporal information.

This difference is important for interpretation. The reported RMSE values reflect the combined effects of the learning algorithm and its assigned feature set. They should not be interpreted as a strictly like-for-like comparison of algorithms using identical predictors. A common-lag robustness analysis would be desirable, but it could not be performed during revision because the original data and estimation code were unavailable. Accordingly, claims about one algorithm being generally superior to another are avoided.

3.5 Treatment of Missing Data

Missing observations are common in large international macroeconomic panels because reporting coverage differs across variables, countries, and years. Removing every observation containing a missing predictor would substantially reduce the sample and could disproportionately exclude countries with weaker statistical reporting systems.

The retained methodology describes a two-stage median-imputation procedure. First, a missing value was replaced with the median of the relevant variable for the corresponding country. Second, when a country did not have sufficient information to calculate a country-specific median, the overall median of that variable was used as a fallback. Median imputation was selected

because it is less sensitive than mean imputation to extreme macroeconomic observations.

The dependent variable was not imputed. Observations without an available GDP-growth outcome or without a valid final feature vector were excluded from model estimation or testing.

In a leakage-free forecasting pipeline, all imputation values should be estimated exclusively from the training period and then applied unchanged to the test period. However, because the original processing code and intermediate data files are unavailable, it was not possible to verify retrospectively whether every imputation statistic was calculated using only the training sample. The study therefore acknowledges possible preprocessing leakage as a reproducibility limitation and does not claim that the imputation procedure has been independently verified.

3.6 Data Preprocessing and Standardization

Continuous predictors were standardized to place variables measured on different scales on a comparable basis. This step is especially relevant for Ridge regression and the Multi-Layer Perceptron because their estimation can be affected by large differences in variable magnitudes.

According to the retained methodological description, the standardization parameters were estimated from the training sample and subsequently applied unchanged to the test sample. This procedure prevents the distribution of the test-period predictors from directly influencing the scaling of the training observations. Tree-based models are less sensitive to differences in scale, but a consistent preprocessing framework was retained across the analysis where applicable.

3.7 Regularized Linear and Tree-Based Machine-Learning Models

Three models are classified as regularized linear or tree-based machine-learning approaches: Ridge regression, Random Forest, and Histogram-Based Gradient Boosting.

Ridge regression is used as the regularized linear benchmark. It reduces coefficient instability by penalizing large coefficient values and is particularly suitable when macroeconomic predictors are correlated. The method trades a limited amount of bias for a potential reduction in forecast variance and provides a comparatively parsimonious benchmark.

Random Forest regression is a tree-based ensemble method that combines predictions from multiple decision trees. Each tree is estimated using a resampled subset of the training observations and a randomized selection of predictors. Aggregating the trees can reduce the instability associated with an individual decision tree while allowing for

nonlinearities and interactions among macroeconomic variables.

Histogram-Based Gradient Boosting regression is a boosting method that constructs decision trees sequentially, with each stage seeking to reduce the prediction errors remaining from earlier stages. The histogram-based implementation groups continuous predictor values into intervals to improve computational efficiency. The model can capture nonlinear relationships and interactions while incorporating controls intended to limit excessive complexity.

The retained manuscript does not provide the complete model settings, including the Ridge regularization value, number of Random Forest trees, tree-depth restrictions, feature-sampling settings, or the full Histogram-Based Gradient Boosting configuration. The original code is unavailable, so these settings cannot be recovered reliably. Rather than introducing assumed values, their absence is reported as a reproducibility limitation.

3.8 AdaBoost and the Neural-Network Model

AdaBoost and the Multi-Layer Perceptron are treated as methodologically distinct approaches.

AdaBoost regression is a classical boosting-based ensemble machine-learning method and is not a deep-learning model. It constructs a sequence of weak prediction models and assigns greater emphasis to observations that were predicted less accurately by preceding models. In this study, the weak learners were decision trees with a maximum depth of three. The retained materials do not report the number of estimators, learning rate, loss specification, or other implementation settings. These details cannot be reconstructed without the original code.

The Multi-Layer Perceptron (MLP) is the neural-network component of the empirical comparison. It is a feed-forward neural network containing two hidden layers with 128 and 64 units, respectively. The model is capable of representing nonlinear relationships and interactions among the lagged macroeconomic predictors. The predictors were standardized before MLP estimation, and early stopping was used to reduce the risk of excessive fitting to the training data.

The retained records do not identify the activation function, optimizer, initial learning rate, batch size, maximum number of training epochs, regularization strength, early-stopping patience, random seed, or complete software environment. These specifications should normally be reported for reproducibility, but they cannot be stated reliably in the absence of the original code and estimation records. The analysis therefore reports only the settings that can be confirmed.

The MLP is used as a neural-network benchmark. Although the literature review discusses recurrent neural networks and Long Short-Term Memory models, no

LSTM or other recurrent architecture was estimated in this study. The empirical findings should consequently not be generalized to all forms of deep learning.

3.9 Model Training and Forecast Evaluation

All models were estimated using observations dated no later than 2021 and evaluated using observations from 2022 and 2023. This common chronological cutoff ensures that each model is assessed on later observations. However, because the model groups use different lag windows, the evaluation is not a fully controlled comparison based on an identical predictor set.

Forecast accuracy is evaluated using root mean squared error (RMSE), calculated from the differences between observed and predicted GDP growth in the test sample. Lower RMSE values indicate better predictive accuracy. RMSE was selected because it assigns relatively greater importance to large forecast errors, which can be particularly consequential during recessions, recoveries, and other periods of macroeconomic disruption. RMSE values are reported to two decimal places throughout the manuscript.

The study compares Ridge regression, Random Forest, Histogram-Based Gradient Boosting, AdaBoost, and the MLP. The retained analysis does not include a historical-mean forecast, a lagged-growth forecast, or another simple forecasting rule. It therefore shows the relative ranking of the five estimated models but does not establish whether they improve upon a naïve macroeconomic benchmark. The absence of such a benchmark is acknowledged as a limitation.

No formal predictive-accuracy test or confidence interval is reported. Although forecast-comparison procedures such as the Diebold–Mariano test are discussed in the literature review, the out-of-sample period contains only two annual forecast dates. This short time dimension does not provide an adequate basis for conventional time-series inference about forecast accuracy. The reported RMSE differences are therefore interpreted descriptively and not as evidence of statistically significant superiority.

The hyperparameters were originally selected as benchmark specifications rather than through exhaustive tuning. However, complete tuning records, software versions, and random-seed information are no longer available. Consequently, the results represent the retained model implementations and should not be interpreted as the best performance theoretically achievable by each model class.

Overall, the methodology combines a broad country-level macroeconomic panel, lagged predictors, median-based missing-data treatment, standardized inputs, and a chronological out-of-sample evaluation. At the same time, differences in the lagged information sets, the absence of a naïve forecasting benchmark, the

short test period, and incomplete reproducibility records limit the strength and generalizability of the comparison. The findings should therefore be understood as descriptive evidence for this particular dataset, predictor set, model implementation, and 2022–2023 evaluation period.

4. Results and Discussion

4.1 Data Characteristics and Exploratory Analysis

Table 1 summarizes the structure and chronological division of the dataset. The full sample contains 2,912 country–year observations covering 213 countries from 2010 to 2023. The training sample contains 2,511 observations from 213 countries for 2010–2021, while the out-of-sample test set contains 401 observations from 206 countries for 2022–2023. Seven countries represented in the full sample are absent from the final test sample because usable observations were unavailable after construction of the test-period variables. The chronological division ensures that later observations are not used to train models intended to forecast earlier outcomes.

Table 1. Dataset Summary and Sample Split (computed by Authors)

Sample	Observations	Countries	Year (Min)	Year (Max)
Full Sample	2,912	213	2010	2023
Training Sample	2,511	213	2010	2021
Test Sample	401	206	2022	2023

Tables 2(a) and 2(b) report descriptive and percentile statistics for GDP growth and the explanatory variables. Average annual country-level GDP growth is 2.85%, with a standard deviation of 6.05 percentage points. The minimum and maximum values of –54.34% and 86.83%, respectively, indicate the presence of severe contractions and strong rebounds in particular country–year observations.

Inflation and real interest rates also contain extreme values, reflecting episodes of substantial macroeconomic instability. The fiscal variables display considerable cross-country dispersion, consistent with differences in fiscal capacity, government size, debt conditions, and reporting practices. GDP per capita is strongly dispersed, reflecting the inclusion of countries at substantially different income levels. Because GDP per capita is measured in current US dollars and was not logarithmically transformed, its extreme upper values should be considered when interpreting the analysis.

Table 2(a). Descriptive Statistics of Key Variables(computed by Authors)

Variable	Mean	Std. Dev.	Min	Max	Variable	25th Percentile	Median	75th Percentile
GDP Growth (% annual)	2.85	6.05	-54.34	86.83	Government Revenue (% of GDP)	19.72	24.21	28.97
Inflation (CPI, %)	5.51	15.92	-4.64	557.20	Tax Revenue (% of GDP)	13.38	16.03	18.71
Unemployment Rate (%)	7.63	5.63	0.10	35.36	Current Account Balance (% of GDP)	-6.81	-2.58	1.69
Real Interest Rate (%)	5.29	7.39	-81.13	61.88	GDP per Capita (USD)	2,300.92	6,789.93	23,279.32
Public Debt (% of GDP)	52.65	24.35	1.85	249.37				
Government Expenditure (% of GDP)	26.43	11.07	0.00	103.73				
Government Revenue (% of GDP)	25.63	15.13	0.00	345.06				
Tax Revenue (% of GDP)	16.45	6.89	0.00	147.64				
Current Account Balance (% of GDP)	-2.19	12.99	-60.88	235.75				
GDP per Capita (USD)	17,809.50	25,463.87	193.01	256,580.52				

Figure 1 presents the distribution of annual country-level GDP growth. The distribution appears asymmetric and heavy-tailed, with both extremely negative and extremely positive observations. These values are consistent with country-specific crises, economic contractions, and subsequent recoveries. The presence of extreme observations makes GDP growth difficult to forecast and is particularly relevant when RMSE is used because this measure assigns greater weight to large forecast errors. The distribution alone, however, does not establish that the underlying relationships between GDP growth and its predictors are nonlinear.

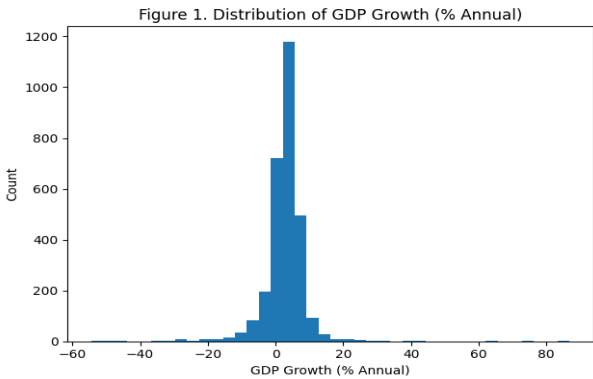


Figure 1. Distribution of Annual GDP Growth (%). (Developed By Authors)

Figure 2 presents the pairwise correlation matrix for the selected macroeconomic variables. Government expenditure, government revenue, and tax revenue display relatively strong positive correlations, indicating potential multicollinearity among the fiscal indicators. This pattern supports the inclusion of Ridge regression, which is designed to stabilize prediction when explanatory variables are correlated.

GDP growth exhibits relatively weak contemporaneous linear correlations with most individual predictors. These weak pairwise correlations

Table 2(b). Percentile Statistics of Key Variables(computed by Authors)

Variable	25th Percentile	Median	75th Percentile
GDP Growth (% annual)	1.00	3.10	5.36
Inflation (CPI, %)	1.50	3.18	5.83
Unemployment Rate (%)	3.98	5.87	9.69
Real Interest Rate (%)	3.25	4.84	6.95
Public Debt (% of GDP)	49.16	49.16	49.16
Government Expenditure (% of GDP)	20.16	25.41	30.68

indicate that no single contemporaneous variable has a strong linear association with GDP growth across the complete sample. They do not, by themselves, demonstrate that GDP growth is generated by nonlinear relationships. Predictive information may instead arise from lagged relationships, combinations of predictors, country heterogeneity, or nonlinear interactions. The nonlinear models are therefore included as alternative empirical specifications rather than being justified solely by the observed correlations.



Figure 2. Correlation Heatmap of Selected Macroeconomic Variables. (Developed By Authors)

Figure 3 shows the unweighted cross-country average of annual GDP growth over time. It is a descriptive average of the country-level observations and should not be interpreted as an expenditure-weighted aggregate global GDP growth rate. The figure shows a pronounced contraction in 2020, followed by a strong recovery. This pattern illustrates the influence of the COVID-19 shock and subsequent reopening on the sample.

The large movements around 2020 and 2021 emphasize the difficulty of forecasting annual GDP growth during periods of synchronized disruption. They also support the use of a chronological evaluation design because a random split could distribute observations from the same crisis and recovery episode across both the training and test samples.

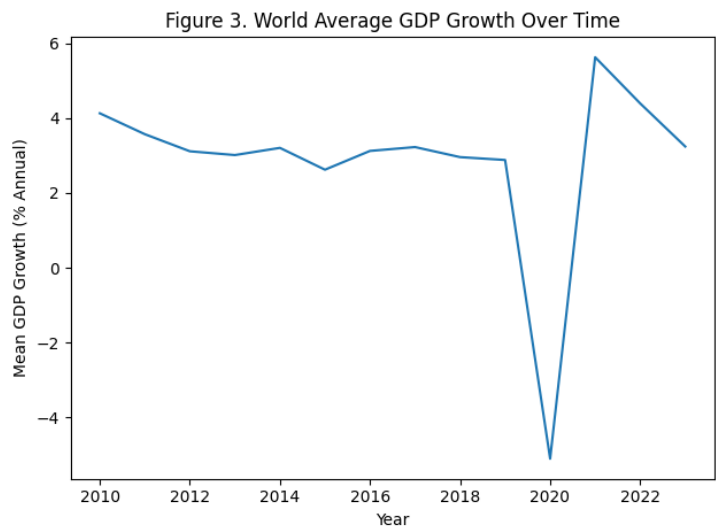


Figure 3. World Average GDP Growth Over Time. (Developed By Authors)

Figure 4 illustrates the cross-country dispersion of annual GDP growth during the most recent years of the sample. The median remains positive in most years, but the dispersion widens considerably during the 2020 contraction. This result indicates that the magnitude of the shock differed substantially across countries. Although dispersion subsequently narrowed, the recovery remained heterogeneous across economies.

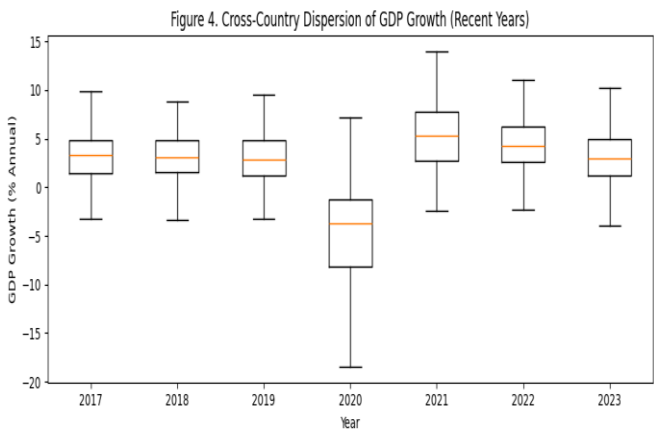


Figure 4. Cross-Country Dispersion of GDP Growth in Recent Years. (Developed By Authors)

The exploratory analysis identifies several features relevant to model selection and interpretation. First, annual GDP growth is volatile and contains extreme observations, making the forecasting problem challenging. Second, the fiscal predictors are correlated, supporting the use of regularization. Third, the weak contemporaneous correlations between GDP growth and

individual predictors indicate that useful forecasting information may depend on lagged variables and combinations of indicators. These observations motivate the comparison of regularized linear, tree-based, boosting, and neural-network approaches. They do not establish in advance which model class should forecast most accurately.

4.2 Machine Learning Baseline Results

Table 3 reports the out-of-sample RMSE values for Ridge regression, Histogram-Based Gradient Boosting, and Random Forest. These three models use the same one-year lag specification and are evaluated on the same 2022–2023 test period. Ridge regression records the lowest observed RMSE at 5.50, followed by Histogram-Based Gradient Boosting at 6.28 and Random Forest at 6.75.

Within this common one-year-lag comparison, the observed RMSE for Ridge regression is 0.78 points lower than that of Histogram-Based Gradient Boosting and 1.25 points lower than that of Random Forest. These values provide a descriptive ranking of the models for the present test sample.

Table 3. Machine Learning Baseline Results (computed by Authors)

Model	RMSE
Ridge Regression	5.50
HistGradientBoosting	6.28
Random Forest	6.75

Figure 5 illustrates the RMSE ranking reported in Table 3. The figure shows that Ridge regression produced the smallest forecast errors among the three models under the present sample, predictor set, and evaluation period. Because no confidence intervals or formal predictive-accuracy tests are available, the figure should not be interpreted as demonstrating statistically significant superiority.

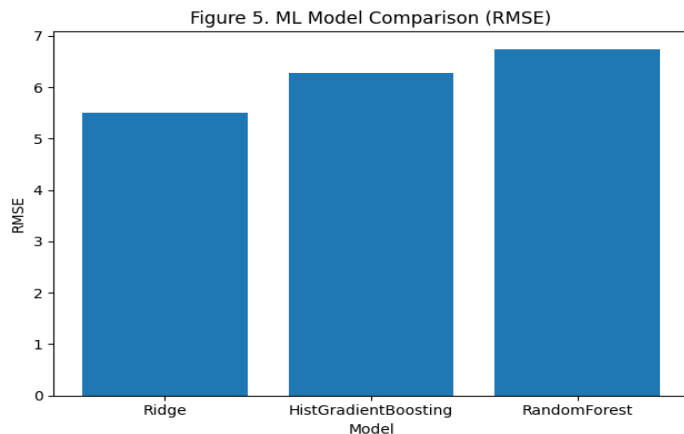


Figure 5. Machine Learning Model Comparison Based on RMSE. (Developed By Authors)

Figure 6 compares observed GDP growth with the predictions produced by Ridge regression. Predictions are concentrated around the middle of the observed distribution, while their correspondence with actual outcomes becomes weaker for unusually large contractions and expansions. This pattern suggests that Ridge regression captures central growth movements more successfully than extreme country–year outcomes.

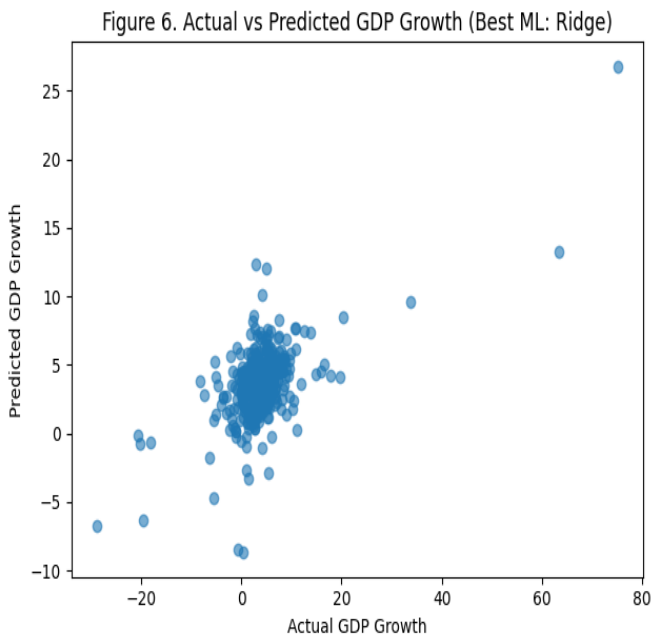


Figure 6. Actual versus Predicted GDP Growth for the Best Machine Learning Model. (Developed By Authors)

Figure 7 presents the distribution of Ridge regression forecast residuals. The residuals appear to be centered close to zero, but the distribution contains relatively large positive and negative errors. These tails are consistent with the extreme GDP-growth observations identified in the exploratory analysis. The result demonstrates the difficulty of predicting major contractions and recoveries using the available annual macroeconomic indicators.

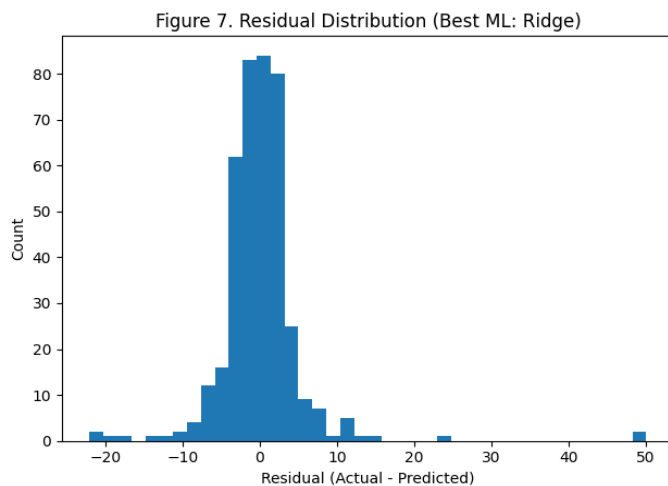


Figure 7. Residual Distribution of the Best Machine Learning Model. (Developed By Authors)

One possible explanation for the relatively favorable performance of Ridge regression is that regularization stabilizes the estimated relationships among correlated macroeconomic predictors. A comparatively parsimonious linear structure may also generalize more reliably when the panel contains many countries but only a limited number of annual periods.

The higher test RMSE values of the two tree-based models do not, by themselves, prove that these models overfitted noise. Establishing overfitting would require training and validation results or additional robustness analysis. Their results show only that, under the retained specifications, the additional flexibility of the tree-based models did not produce lower test-period RMSE than Ridge regression.

Overall, the results in this subsection show that Ridge regression provides the lowest observed RMSE among the models using the common one-year lag specification. This conclusion is specific to the retained dataset, predictors, preprocessing decisions, model settings, and 2022–2023 test period.

4.3 Deep Learning–Style Model Results

Table 4 reports the results for AdaBoost and the Multi-Layer Perceptron. AdaBoost is a classical boosting-based ensemble machine-learning method, whereas the MLP is the neural-network model included in the analysis. AdaBoost should therefore not be classified as a deep-learning model.

AdaBoost, using decision trees with a maximum depth of three, records an RMSE of 7.13. The MLP, with hidden layers containing 128 and 64 units, records an RMSE of 8.12. Thus, AdaBoost has a test RMSE that is 0.99 points lower than that of the MLP. This difference provides a descriptive comparison between the two retained implementations and should not be interpreted

as evidence about boosting and neural networks in general.

Table 4. Deep Learning–Style Model Results (computed by Authors)

Model	RMSE
AdaBoost (tree depth = 3)	7.13
MLP (128, 64)	8.12

Figure 8 compares the observed GDP-growth outcomes with the AdaBoost predictions. The predictions are concentrated around the central part of the observed distribution, while the model has greater difficulty reproducing extreme contractions and expansions. This pattern is consistent with the challenge of forecasting infrequent macroeconomic shocks from annual lagged indicators.

Figure 8 (DL). Actual vs Predicted GDP Growth (Best DL: DL_Model_A_MLP(128,64))

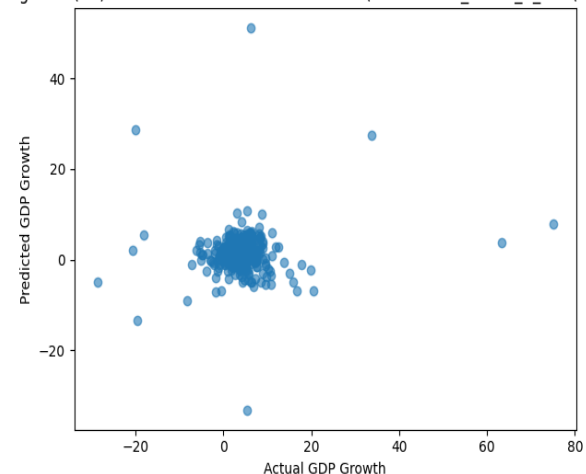


Figure 8. Actual versus Predicted GDP Growth for the Best Deep Learning–Style Model. (Developed By Authors)

Figure 9 presents the distribution of AdaBoost forecast residuals. The residuals appear to be centered close to zero but display relatively heavy tails, indicating the presence of several large forecast errors. As with Ridge regression, the largest errors are likely associated with observations that differ substantially from the more common growth outcomes in the sample.

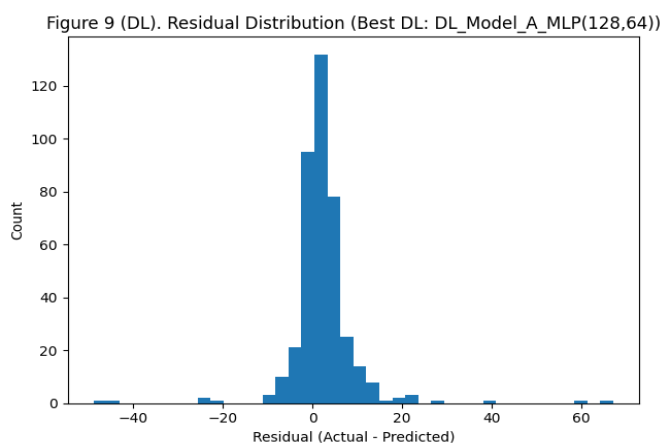


Figure 9. Residual Distribution of the Best Deep Learning–Style Model. (Developed By Authors)

The lower observed RMSE of AdaBoost relative to the MLP may reflect differences in how the two models respond to the available sample structure. AdaBoost combines shallow trees sequentially, whereas the MLP estimates a more highly parameterized nonlinear mapping. The panel contains a relatively large number of country–year observations but only a short annual time dimension. This structure may limit the MLP’s ability to learn stable temporal relationships.

Nevertheless, the available evidence does not demonstrate that the MLP overfitted the training data because training and validation losses are not reported. Differences in hyperparameters, optimization, regularization, and model implementation could also contribute to its observed performance. The reported result therefore concerns the particular MLP specification used in this study and does not establish that neural networks generally perform poorly in macroeconomic forecasting.

Across the five evaluated models, the observed RMSE ranking is Ridge regression at 5.50, Histogram-Based Gradient Boosting at 6.28, Random Forest at 6.75, AdaBoost at 7.13, and the MLP at 8.12. Ridge regression therefore records the lowest observed test RMSE, while the MLP records the highest.

This overall ranking must be interpreted cautiously because the models do not all use identical information sets. Ridge regression, Random Forest, and Histogram-Based Gradient Boosting use one-year lags, whereas AdaBoost and the MLP use lag windows of one to three years. Differences in RMSE across these groups consequently reflect both the modeling algorithms and their respective feature sets. The comparison does not isolate the effect of algorithm choice alone.

The forecast evaluation also does not include a historical-mean forecast, lagged-growth forecast, or another naïve benchmark. It is therefore possible to compare the five retained models with one another, but it is not possible to determine whether they improve upon a simple macroeconomic forecasting rule. This

omission should be addressed in future work when the data and code are available.

Formal uncertainty estimates and predictive-accuracy tests are not reported. The test period covers only two annual forecast dates, 2022 and 2023, even though it contains multiple country observations. This short time dimension makes conventional forecast-comparison tests difficult to justify. Accordingly, terms such as “lowest observed RMSE” and “better descriptive performance” are used instead of claims of statistically significant superiority.

Taken together, the results indicate that greater model complexity did not correspond to lower forecast errors in this particular application. A possible interpretation is that regularization and parsimony are beneficial when annual macroeconomic relationships are noisy and the time dimension is short. However, the findings do not demonstrate that Ridge regression generally dominates tree-based, boosting, or neural-network methods. They apply only to the present country-level panel, predictor set, lag specifications, retained model implementations, and 2022–2023 evaluation period.

5. Conclusion and Future Research

This study compared regularized linear, tree-based, boosting, and neural-network approaches for forecasting annual country-level GDP growth. The analysis used an unbalanced macroeconomic panel covering 213 countries from 2010 to 2023. To preserve the forecasting sequence, observations from 2010–2021 were used for training, while observations from 2022–2023 were reserved for out-of-sample evaluation.

Among the five evaluated models, Ridge regression recorded the lowest observed out-of-sample RMSE at 5.50. It was followed by Histogram-Based Gradient Boosting at 6.28, Random Forest at 6.75, AdaBoost at 7.13, and the Multi-Layer Perceptron at 8.12. These results indicate that greater model complexity did not correspond to lower forecast errors in this particular application. The comparatively favorable performance of Ridge regression is consistent with the potential value of regularization and model parsimony when macroeconomic predictors are correlated, annual time-series histories are short, and economic outcomes are affected by large and infrequent shocks.

The study also clarifies the methodological classification of the models. Ridge regression is a regularized linear model; Random Forest, Histogram-Based Gradient Boosting, and AdaBoost are conventional ensemble machine-

learning methods; and the MLP is the neural-network component of the comparison. In particular, AdaBoost is not classified as a deep-learning method. No recurrent neural network or Long Short-Term Memory model was estimated.

The findings should be interpreted descriptively rather than as evidence of statistically significant superiority. The evaluation period contains only two annual forecast dates, making conventional time-series tests of predictive accuracy difficult to justify. Furthermore, Ridge regression, Random Forest, and Histogram-Based Gradient Boosting used one-year lags, whereas AdaBoost and the MLP used lag windows of one to three years. Comparisons across these groups therefore reflect differences in both the learning algorithms and their information sets. The study does not isolate the effect of model choice under a completely identical feature specification.

Several additional limitations should be recognized. First, the evaluation does not include a historical-mean forecast, a lagged-GDP-growth forecast, or another naïve benchmark. The reported results consequently establish a ranking among the five evaluated models but do not demonstrate that these models improve upon simple forecasting rules. Second, the original data-processing and estimation code were unavailable during revision. Complete data-provenance records, preprocessing verification, software versions, random seeds, and several model hyperparameters could therefore not be reconstructed. Third, the use of country-specific and overall median imputation may reduce genuine variation in variables with substantial missingness. It was also not possible to verify retrospectively whether every imputation statistic had been estimated exclusively from the training period. Fourth, GDP per capita was measured in current US dollars and was not logarithmically transformed despite its highly skewed distribution. Finally, the sample includes the unusual COVID-19 contraction and recovery, which may have had a substantial influence on the model rankings.

Future research should re-estimate all models using an identical lagged information set. A common-lag specification would provide a more controlled comparison of the forecasting algorithms, while separate lag-window robustness checks could evaluate the value of additional temporal information. Future studies should also include simple benchmarks, such as the historical mean and lagged GDP growth, to determine

whether more complex methods provide incremental forecasting value.

A longer evaluation period with rolling or expanding-window forecasts would permit a more reliable assessment of performance across different economic conditions. With a sufficient number of forecast dates, future research could report confidence intervals, forecast-error uncertainty, and appropriate tests of predictive accuracy. Performance could also be examined separately across income groups, geographical regions, crisis and non-crisis periods, and countries with different levels of data availability.

Data preparation could be strengthened by obtaining all indicators directly from their original providers, reporting the relevant indicator codes and extraction dates, retaining a transparent duplicate-resolution procedure, and estimating all imputation and scaling parameters exclusively from each training window. Robustness analyses could consider logarithmic GDP per capita, alternative missing-data methods, outlier-sensitive and outlier-robust evaluation measures, and additional predictors representing institutional, financial, trade, and political conditions.

Future neural-network research should report complete architectural, optimization, regularization, and early-stopping settings. Recurrent or sequence-based models, including LSTM networks, could be examined if longer time series or higher-frequency data become available. Their inclusion should be based on an appropriate data structure rather than an assumption that greater architectural complexity necessarily produces better forecasts.

Overall, the study provides evidence that a comparatively simple regularized linear model can produce competitive country-level annual GDP-growth forecasts under a disciplined chronological evaluation. However, the result is specific to the present panel, predictor set, feature construction, model implementations, and 2022–2023 test period. It should not be interpreted as demonstrating that Ridge regression generally dominates ensemble learning or neural networks in macroeconomic forecasting.

Data Availability Statement

Publicly available datasets were analyzed in this study. The data can be found from World Bank Open Data / World Development Indicators and the Kaggle dataset titled “**Global Economic Indicators (2010–2025)**”.

Dataset link:

<https://www.kaggle.com/datasets/tanishksharma9905/global-economic-indicators-20102025>

Funding

Funding information is not available.

Acknowledgements

The authors would like to thank the World Bank Open Data platform for providing open-access macroeconomic indicators. The authors also acknowledge the open-source Python community and related data analysis libraries that supported the empirical analysis and visualization in this study.

Institutional Review Board Statement

Not Applicable.

Informed Consent Statement

Not Applicable.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

References

- [1] STOCK J H, WATSON M W. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 2002, 20(2): 147–162. DOI: 10.1198/073500102317351921.
- [2] GIANNONE D, REICHLIN L, SMALL D. Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 2008, 55(4): 665–676. DOI: 10.1016/j.jmoneco.2008.05.010.
- [3] BAI J, NG S. Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 2002, 70(1): 191–221. DOI: 10.1111/1468-0262.00273.
- [4] DE MOL C, GIANNONE D, REICHLIN L. Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? *Journal of Econometrics*, 2008, 146(2): 318–328. DOI: 10.1016/j.jeconom.2008.08.011.
- [5] BANBURA M, GIANNONE D, REICHLIN L. Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics*, 2010, 25(1): 71–92. DOI: 10.1002/jae.1137.
- [6] CARRIERO A, CLARK T E, MARCELLINO M. Bayesian VARs: Specification Choices and Forecast Accuracy. *Journal of Applied Econometrics*, 2015, 30(1): 46–73. DOI: 10.1002/jae.2315.
- [7] KOOP G, KOROBILIS D. Forecasting Inflation Using Dynamic Model Averaging. *International Economic Review*, 2012, 53(3): 867–886. DOI: 10.1111/j.1468-2354.2012.00704.x.
- [8] SCHORFHEIDE F, SONG D. Real-Time Forecasting With a Mixed-Frequency VAR. *Journal of Business & Economic Statistics*, 2015, 33(3): 366–380. DOI: 10.1080/07350015.2014.954707.
- [9] DIEBOLD F X, MARIANO R S. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 1995, 13(3): 253–263. DOI: 10.1080/07350015.1995.10524599.
- [10] WEST K D. Asymptotic Inference About Predictive Ability. *Econometrica*, 1996, 64(5): 1067–1084. DOI: 10.2307/2171956.
- [11] VARIAN H R. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 2014, 28(2): 3–28. DOI: 10.1257/jep.28.2.3.
- [12] EINAV L, LEVIN J. Economics in the Age of Big Data. *Science*, 2014, 346(6210): 1243089. DOI: 10.1126/science.1243089.
- [13] MULLAINATHAN S, SPIESS J. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 2017, 31(2): 87–106. DOI: 10.1257/jep.31.2.87.
- [14] BELLONI A, CHERNOZHUKOV V, HANSEN C. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 2014, 28(2): 29–50. DOI: 10.1257/jep.28.2.29.
- [15] ATHEY S, IMBENS G W. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 2019, 11: 685–725. DOI: 10.1146/annurev-economics-080217-053433.
- [16] HOERL A E, KENNARD R W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970, 12(1): 55–67. DOI: 10.1080/00401706.1970.10488634.
- [17] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 1996, 58(1): 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
- [18] BREIMAN L. Random Forests. *Machine Learning*, 2001, 45: 5–32. DOI: 10.1023/A:1010933404324.
- [19] BIAU G, SCORNET E. A Random Forest Guided Tour. *TEST*, 2016, 25: 197–227. DOI: 10.1007/s11749-016-0481-7.
- [20] FRIEDMAN J H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 2001, 29(5): 1189–1232. DOI: 10.1214/aos/1013203451.
- [21] FREUND Y, SCHAPIRE R E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997, 55(1): 119–139. DOI: 10.1006/jcss.1997.1504.
- [22] ANG A, PIAZZESI M, WEI M. What Does the Yield Curve Tell Us About GDP Growth? *Journal of Econometrics*, 2006, 131(1–2): 359–403. DOI: 10.1016/j.jeconom.2005.01.032.

- [23] MARCELLINO M, STOCK J H, WATSON M W. A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. *Journal of Econometrics*, 2006, 135(1-2): 499-526. DOI: 10.1016/j.jeconom.2005.07.020.
- [24] HYNDMAN R J, KOEHLER A B. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 2006, 22(4): 679-688. DOI: 10.1016/j.ijforecast.2006.03.001.
- [25] MAKRIDAKIS S, SPILIOTIS E, ASSIMAKOPOULOS V. The M4 Competition: 100,000 Time Series and 61 Forecasting Methods. *International Journal of Forecasting*, 2020, 36(1): 54-74. DOI: 10.1016/j.ijforecast.2019.04.014.
- [26] ZHANG G, PATUWO B E, HU M Y. Forecasting With Artificial Neural Networks: The State of the Art. *International Journal of Forecasting*, 1998, 14(1): 35-62. DOI: 10.1016/S0169-2070(97)00044-7.
- [27] HORNIK K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 1991, 4(2): 251-257. DOI: 10.1016/0893-6080(91)90009-T.
- [28] LECUN Y, BENGIO Y, HINTON G. Deep Learning. *Nature*, 2015, 521: 436-444. DOI: 10.1038/nature14539.
- [29] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory. *Neural Computation*, 1997, 9(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.
- [30] MEDEIROS M C, VASCONCELOS G F R, VEIGA A, ZILBERMAN E. Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, 2021, 39(1): 98-119. DOI: 10.1080/07350015.2019.1637745.
- [31] GOULET COULOMBE P, LEROUX M, STEVANOVIC D, SURPRENANT S. How Is Machine Learning Useful for Macroeconomic Forecasting? *Journal of Applied Econometrics*, 2022, 37(5): 920-964. DOI: 10.1002/jae.2910.
- [32] GIACOMINI R, WHITE H. Tests of Conditional Predictive Ability. *Econometrica*, 2006, 74(6): 1545-1578. DOI: 10.1111/j.1468-0262.2006.00718.x.
- [33] CLARK T E, WEST K D. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *Journal of Econometrics*, 2007, 138(1): 291-311. DOI: 10.1016/j.jeconom.2006.05.023.
- [34] MCCracken M W, NG S. FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 2016, 34(4): 574-589. DOI: 10.1080/07350015.2015.1086655.
- [35] PESARAN M H, SCHUERMANN T, WEINER S M. Modeling Regional Interdependencies Using a Global Error-Correcting Macroeconometric Model. *Journal of Business & Economic Statistics*, 2004, 22(2): 129-162. DOI: 10.1198/073500104000000019.

参考文献:

- [1] STOCK J H, WATSON M W. 使用扩散指数进行宏观经济预测. *商业与经济统计学杂志*, 2002, 20(2): 147-162. DOI: 10.1198/073500102317351921.
- [2] GIANNONE D, REICHLIN L, SMALL D. 即时预测: 宏观经济数据的实时信息含量. *货币经济学杂志*, 2008, 55(4): 665-676. DOI: 10.1016/j.jmoneco.2008.05.010.
- [3] BAI J, NG S. 近似因子模型中因子数量的确定. *计量经济学*, 2002, 70(1): 191-221. DOI: 10.1111/1468-0262.00273.
- [4] DE MOL C, GIANNONE D, REICHLIN L. 使用大量预测变量进行预测: 贝叶斯收缩是否是主成分方法的有效替代方案? *计量经济学杂志*, 2008, 146(2): 318-328. DOI: 10.1016/j.jeconom.2008.08.011.
- [5] BANBURA M, GIANNONE D, REICHLIN L. 大型贝叶斯向量自回归模型. *应用计量经济学杂志*, 2010, 25(1): 71-92. DOI: 10.1002/jae.1137.
- [6] CARRIERO A, CLARK T E, MARCELLINO M. 贝叶斯VAR模型: 设定选择与预测精度. *应用计量经济学杂志*, 2015, 30(1): 46-73. DOI: 10.1002/jae.2315.
- [7] KOOP G, KOROBILIS D. 使用动态模型平均法预测通货膨胀. *国际经济评论*, 2012, 53(3): 867-886. DOI: 10.1111/j.1468-2354.2012.00704.x.
- [8] SCHORFHEIDE F, SONG D. 基于混合频率VAR的实时预测. *商业与经济统计学杂志*, 2015, 33(3): 366-380. DOI: 10.1080/07350015.2014.954707.
- [9] DIEBOLD F X, MARIANO R S. 预测精度比较. *商业与经济统计学杂志*, 1995, 13(3): 253-263. DOI: 10.1080/07350015.1995.10524599.
- [10] WEST K D. 关于预测能力的渐近推断. *计量经济学*, 1996, 64(5): 1067-1084. DOI: 10.2307/2171956.
- [11] VARIAN H R. 大数据: 计量经济学的新方法. *经济展望杂志*, 2014, 28(2): 3-28. DOI: 10.1257/jep.28.2.3.
- [12] EINAV L, LEVIN J. 大数据时代的经济学. *科学*, 2014, 346(6210): 1243089. DOI: 10.1126/science.1243089.
- [13] MULLAINATHAN S, SPIESS J. 机器学习: 一种应用计量经济学方法. *经济展望杂志*, 2017, 31(2): 87-106. DOI: 10.1257/jep.31.2.87.
- [14] BELLONI A, CHERNOZHUKOV V, HANSEN C. 高维方法以及结构效应和处理效应推断. *经济展望杂志*, 2014, 28(2): 29-50. DOI: 10.1257/jep.28.2.29.
- [15] ATHEY S, IMBENS G W. 经济学家应了解的机器学习方法. *经济学年度评论*, 2019, 11: 685-725. DOI: 10.1146/annurev-economics-080217-053433.
- [16] HOERL A E, KENNARD R W. 岭回归: 非正交问题的有偏估计. *技术计量学*, 1970, 12(1): 55-67. DOI: 10.1080/00401706.1970.10488634.

- [17] TIBSHIRANI R. 通过Lasso进行回归收缩与变量选择. *皇家统计学会杂志B辑*, 1996, 58(1): 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
- [18] BREIMAN L. 随机森林. *机器学习*, 2001, 45: 5–32. DOI: 10.1023/A:1010933404324.
- [19] BIAU G, SCORNET E. 随机森林方法导论. *TEST*, 2016, 25: 197–227. DOI: 10.1007/s11749-016-0481-7.
- [20] FRIEDMAN J H. 贪婪函数逼近: 梯度提升机. *统计年鉴*, 2001, 29(5): 1189–1232. DOI: 10.1214/aos/1013203451.
- [21] FREUND Y, SCHAPIRE R E. 在线学习的决策理论推广及其在Boosting中的应用. *计算机与系统科学杂志*, 1997, 55(1): 119–139. DOI: 10.1006/jcss.1997.1504.
- [22] ANG A, PIAZZESI M, WEI M. 收益率曲线能告诉我们什么关于GDP增长的信息? *计量经济学杂志*, 2006, 131(1–2): 359–403. DOI: 10.1016/j.jeconom.2005.01.032.
- [23] MARCELLINO M, STOCK J H, WATSON M W. 宏观经济时间序列预测中直接法与迭代多步AR方法的比较. *计量经济学杂志*, 2006, 135(1–2): 499–526. DOI: 10.1016/j.jeconom.2005.07.020.
- [24] HYNDMAN R J, KOEHLER A B. 对预测精度度量方法的再考察. *国际预测杂志*, 2006, 22(4): 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.
- [25] MAKRIDAKIS S, SPILIOTIS E, ASSIMAKOPOULOS V. M4竞赛: 100000个时间序列与61种预测方法. *国际预测杂志*, 2020, 36(1): 54–74. DOI: 10.1016/j.ijforecast.2019.04.014.
- [26] ZHANG G, PATUWO B E, HU M Y. 使用人工神经网络进行预测: 研究现状. *国际预测杂志*, 1998, 14(1): 35–62. DOI: 10.1016/S0169-2070(97)00044-7.
- [27] HORNIK K. 多层前馈网络的逼近能力. *神经网络*, 1991, 4(2): 251–257. DOI: 10.1016/0893-6080(91)90009-T.
- [28] LECUN Y, BENGIO Y, HINTON G. 深度学习. *自然*, 2015, 521: 436–444. DOI: 10.1038/nature14539.
- [29] HOCHREITER S, SCHMIDHUBER J. 长短期记忆网络. *神经计算*, 1997, 9(8): 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
- [30] MEDEIROS M C, VASCONCELOS G F R, VEIGA Á, ZILBERMAN E. 在数据丰富环境中预测通货膨胀: 机器学习方法的优势. *商业与经济统计学杂志*, 2021, 39(1): 98–119. DOI: 10.1080/07350015.2019.1637745.
- [31] GOULET COULOMBE P, LEROUX M, STEVANOVIC D, SURPRENANT S. 机器学习如何有助于宏观经济预测? *应用计量经济学杂志*, 2022, 37(5): 920–964. DOI: 10.1002/jae.2910.
- [32] GIACOMINI R, WHITE H. 条件预测能力检验. *计量经济学*, 2006, 74(6): 1545–1578. DOI: 10.1111/j.1468-0262.2006.00718.x.
- [33] CLARK T E, WEST K D. 嵌套模型中相等预测精度的近似正态检验. *计量经济学杂志*, 2007, 138(1): 291–311. DOI: 10.1016/j.jeconom.2006.05.023.
- [34] MCCracken M W, NG S. FRED-MD: 用于宏观经济研究的月度数据库. *商业与经济统计学杂志*, 2016, 34(4): 574–589. DOI: 10.1080/07350015.2015.1086655.
- [35] PESARAN M H, SCHUERMANN T, WEINER S M. 使用全球误差修正宏观计量经济模型刻画区域相互依赖关系. *商业与经济统计学杂志*, 2004, 22(2): 129–162. DOI: 10.1198/073500104000000019.

Manuscript Information

Word count: 10,371 words (excluding references).

Peer-Review Record

Fast-track status: Not fast-tracked.

First-round reviews received: 3 reports.

Revision cycles completed: 3 rounds.

Final version submitted: October 8, 2026

Disclaimer / Publisher's Note

The statements, opinions, and data contained in this article are solely those of the authors and do not necessarily represent the views of the *Journal of Hunan University (Natural Sciences)* or its editorial team. The journal and its editors disclaim any responsibility for injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content of this article.