JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

PREDICTING ANNUAL GDP GROWTH ACROSS COUNTRIES USING MACROECONOMIC PANEL DATA

Zeeshan Shakoor[1], Muazzam Ali[*2], Abdul Manan[3], M. Usman Hashmi[4], Affan Ahmad[5]

[1]*Faculty of Sciences, Superior University, Lahore, Pakistan*
[*2,3,4,5]*Faculty of Computer Science and Information Technology, Superior University, Lahore, Pakistan*

[*2]muazzamali@superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.22828222

Keywords
GDP growth; machine learning; panel data; out-of-time evaluation; directional accuracy; explainable AI

Article History
Received: 11 June 2026
Accepted: 02 September 2026
Published: 18 September 2026

Corresponding Author: *
Muazzam Ali

Abstract

Forecasting annual GDP growth across a large set of economies remains difficult because the data combine structural differences, uneven reporting and occasional shocks of exceptional size. We compare linear regression, random forest, XGBoost, LightGBM and CatBoost using 3,472 country-year records for 217 economies and territories over 2010-2025. The workflow includes missingness indicators, median imputation, winsorisation, robust scaling, lagged variables and country and year encodings, followed by a chronological holdout. All five models produce a negative out-of-time R^2: LightGBM is closest to the holdout-mean benchmark at -0.0076, whereas linear regression records -0.7106. The tree ensembles nonetheless improve markedly on the linear model. LightGBM reduces RMSE by 23.25%, CatBoost reduces MAE by 44.37%, and random forest achieves the highest reported sign agreement at 95.08%. That sign result should be interpreted cautiously. The available outputs do not include the class balance, a majority-sign or persistence benchmark, paired forecast errors, rolling-origin estimates or country-clustered uncertainty. Model explanations are similarly sensitive to the diagnostic used: split importance gives greatest weight to lagged GDP growth, while mean absolute SHAP values emphasise GDP-per-capita variables and the year code. The evidence therefore favours nonlinear ensembles over the fitted linear specification, but it does not establish superior point forecasting, a validated recession-warning rule or causal macroeconomic effects.

1.0 Introduction

Annual GDP growth enters fiscal plans, monetary-policy discussions, sovereign-risk assessments and private investment decisions. Yet it is a difficult quantity to predict across countries. National accounts are revised, publication calendars differ, and crises can break relationships that appeared stable in earlier years. Structural models remain useful because their assumptions are explicit [1,2], while forecast pooling can protect against some forms of model instability [3]. Data-rich nowcasting and time-varying parameter models address other parts of the problem, although neither eliminates sensitivity to the information set and evaluation period [4,5]. Machine learning offers a different compromise between structure and flexibility. Random forests average many decorrelated trees [6,7], and boosting methods build nonlinear predictors

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

sequentially [8]. XGBoost, LightGBM and CatBoost extend that idea with efficient optimisation and, in CatBoost's case, specialised treatment of categorical information [9-11]. Such flexibility is useful only when the validation design resembles the intended application. Chronological separation and forecast-origin information sets are indispensable; otherwise, revised or contemporaneous variables can make a retrospective exercise look like a genuine forecast [12,13]. These design choices matter especially in policy settings, where models are often used to complement rather than replace established analysis [14]. Three concerns are central here. Several regressors are measured in the same year as GDP growth, so the exercise is better described as out-of-time prediction or retrospective nowcasting unless their release dates are shown to precede the target. A high sign-hit rate may also arise when expansions dominate the holdout. Finally, feature-attribution plots describe a fitted function; without additional assumptions, they do not identify structural economic effects [15-17].

Against that background, the paper compares four tree ensembles with linear regression on a common chronological holdout. It examines whether point-error and sign metrics lead to the same model choice and re-reads the published explanation plots without assigning them a causal meaning. The revision also extracts a small set of additional, fully reproducible quantities from the reported table: approximate sign counts, descriptive Wilson intervals, the implied mean-benchmark RMSE and error reductions relative to linear regression. No model is retrained for these calculations.

2.0 Related Work

Large macroeconomic databases have led to frequent employment of a relatively new range of techniques involving regularization and nonlinear forecasting [18-20]. These data sets enable the juxtaposition of both fiscal and financial indicators, as well as to demographic and institutional indicators, but also raise the possibility of overfitting and unstable variable selection. Therefore, regularization is of great importance as it is used to reduce the complexity of the model and prevents the impact of insignificant/exploratory predictors. However, as noted in the literature, machine learning should not be seen as a single tool but also as the thought that the more complex the model, the better the predictions will be [21]. Algorithms have varying assumptions, noise sensitivities and are of different representational capacity for interactions. In some cases, a dense predictor set can also be perceived as a sparse predictor set given a specific sample. A dense predictor set can also seem like a sparse predictor set in a specific sample even though the underlying economic signal exists, but isn't common across multiple related predictors. [22] Therefore, by choosing one factor and leaving out a second factor, it is not concluded that the second is an economically unimportant factor. The chosen variable can only be considered as reflecting a wider (macro-economic) process.

This current result also replicates what macro-economic applications suggest: most often the fact that improvements in the forecast come not from one particular software library or algorithm, but because the result should be in the form of nonlinearities and proper transformations and regularization procedures [23]. The links of these properties to GDP growth are that effects of investment, government expenditures and financial conditions can vary with economic states. Nonlinearities and interactions can be modeled by random forests without a need to specify them ahead, which has led to adaptation to macroeconomic forecasting [24]. But the more flexible they are, the greater the risk of needing to properly validate it as they may learn shortcuts and/or some permanent country differences. Model rankings can also be altered depending on the forecast horizon and the transformation and test period of the inflation data [25]. The model that works well when the economic situation is stable may not be as effective in times of crisis and big policy shifts. The use of predictive superiority must hence be qualified not to suggest general statements, but

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

to indicate the sample design and the evaluation time.

Observations cannot be simply lumped together from different cross-country panels. What measurements are taken from the same economy are also serially related and economies in the same year may be impacted by similar global events. This structure can be included in country and year terms, but note that integer country codes in a tree model are not the same as econometric fixed effects [26]. As a data structure of these codes is used as a predictor in tree-based models, it can create economically uninterpretable numeric partitions. They should therefore be seen as predictive rather than as structure, and they don't necessarily have to be liked. Evaluation should also be based on improvement, not just on improvement compared to a fitted linear model, but on improvement over simple heuristic models like the historical mean or persistence/prevailing sign. A more sophisticated model could fare better than a regression, but not substantially gain any advantage over these practical standards. Forecast tests, where the forecast loss's difference is measured on both directions, can be used to check systematic forecast loss differences and direction tests can guess the direction of the forecast loss more successfully than the benchmark [27,28]. It is better to take into account both the amount and the direction of the errors when judging the performance of a forecast.

SHAP decomposes an individual prediction into the sum of the contributions of each feature in a tree ensemble and can be computed efficiently [29,30]. It can find variables that are most important to the predictions that are made and indicate how much that variable contributes to those predictions from one observation to the next. Partial dependence, on the other hand, refers to the average fitted response when one predictor is varied. Both approaches are descriptive in nature and should be used with due caution in some instances where the predictors are correlated. Redistribution of SHAP contributions across similar attributes and evaluations of unrealistic combinations of attributes. The effects are more likely to be stable in accumulated local effects since they consider only changes that have taken place in the areas for which predictors have been observed. All these procedures are observational methods that do not prevent any cause/effect. For a causal interpretation, assumptions need to be made about the absence of confounding, the simultaneity effect, and identification—assumptions not provided by the prediction model itself [17].

3.0 Methodology

3.1 Research design and scope

We frame the analysis as a supervised regression comparison with a chronological holdout. Each row represents an economy in a calendar year, and every model is fitted to the same engineered feature matrix. Because the predictor list contains same-year macroeconomic variables and no release-vintage table is available, the empirical target is out-of-time predictive association rather than a verified one-year-ahead real-time forecast.

3.2 Data source and analytical sample

The processed file, world_bank_data_2025.csv, is described as an extract from World Development Indicators. It contains 3,472 rows: 217 economies and territories observed on a 16-year grid from 2010 through 2025. Since 217 x 16 equals 3,472, the row grid is balanced; missingness occurs within indicator fields rather than through absent country-year rows. GDP growth (annual %) is the response. Fifteen initial predictors cover nominal output and income, GDP per capita, consumer-price and GDP-deflator inflation, public debt, government revenue and expense, the current-account balance and unemployment. The World Bank GDP-growth series combines national and international statistical sources and is distributed under a CC BY 4.0 licence [31].

3.3 Preprocessing and feature construction

Preprocessing proceeds in a fixed sequence. Variables with more than 50% missing values

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

are removed, after which missingness indicators are created for retained fields. Country medians provide the first imputation level and a global median is used when an economy has no observed value for a feature. Continuous predictors are winsorised at the 1st and 99th percentiles, highly skewed size variables receive a log1p transformation, and selected indicators are lagged by one year. GDP growth is also lagged. Country and year codes complete the feature set. The modelling pipeline applies median imputation and RobustScaler to the resulting matrix.

A valid replication must estimate medians, winsorisation cut-offs, scaling parameters and any data-driven transformation choices on the training period alone, then apply them unchanged to the holdout. The material available for this review does not contain an execution log that confirms the fitting scope of each transformation. The results should therefore be read with this reproducibility limitation in mind.

3.4 Temporal split and information set

Observations are ordered by year and split at the 80th percentile of the timeline. The panel dimensions and reported percentages are consistent with a final three-year block of 651 rows and an earlier block of 2,821 rows, although the exact row-level split should be confirmed from code. The reported evaluation uses this single holdout rather than random k-fold resampling. Consequently, the estimates describe one recent period and do not reveal how rankings change across alternative forecast origins.

3.5 Models

Five regressors are compared using fixed specifications drawn from the reported analysis:

- Linear regression: an unregularised parametric baseline.
- Random forest: 500 trees, maximum depth 8, minimum 5 observations per leaf and $random\_state = 42$. [6,7]
- XGBoost: 500 estimators, learning rate 0.03, maximum depth 4, row subsampling 0.80, column subsampling 0.80 and $random\_state = 42$. [9]
- LightGBM: 500 estimators, learning rate 0.03, maximum depth 4 and $random\_state = 42$. [10]
- CatBoost: 500 iterations, depth 6, learning rate 0.03 and $random\_seed = 42$. [11]

No training-only tuning protocol is reported, so the settings are treated as fixed model specifications rather than as the outcome of an unbiased hyperparameter search. Country and year codes are predictive encodings; they should not be interpreted as evidence that econometric fixed effects have been identified.

3.6 Evaluation metrics

The accuracy of the points is expressed in RMSE, MAE and R^2. Here R2 measures squared error against prediction using the 'holdout mean', which if negative will indicate that the "fitted" model does worse. The MAPE is retained to represent the original analysis and is represented as a percentage, not a ratio. Since the annual gdp growth could be null, near null or even negative, MAPE may be unstable and should not be used alone to rank the models [32]. The accuracy of direction shows percentage of the rows where the direction of the prediction growth rate and the actual growth rate were the same.

3.7 Secondary calculations from reported outputs

The secondary analysis is based only on Table 1 and an assumed number of holdouts of 651. Sign percentages are reported in a roundabout fashion for the nearest possible integer count, then mapped to Wilson intervals, which are then compared to linear regression to derive RMSE, MAE and directional-error reductions. These intervals are descriptive in nature, and do not take into account dependence within countries and within years. A small inconsistancy is preserved transparently: 93.54%, for CatBoost maps it's about 609 correct signs, but 609/651 is rounded to 93.55%. The result indicates truncation which could be due to a different effective denominator, or to a reporting-rounding

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

problem in the source output. This would only be resolved at an observation level. These comparisons are not an experiment to remove features but one of a "model family" comparison.

3.8 Explainability analysis

The LightGBM diagnostics comprise split-based importance, mean absolute SHAP values and partial-dependence curves for lagged GDP growth and the current-account balance. These measures will address different questions and do not necessarily have to produce the same ranking. Therefore we use the terms views of the fitted model, and not elasticities or causal effects.

4.0 Results

4.1 Comparative performance of machine learning models

All models' predictive performance presented in Table 1. The analysis of the results shows that the nonlinear ensemble models in general give better results than the linear regression models when tested with all metrics. Overall, of the tested solutions, LightGBM showed the best performance in the squared-error category with the smallest RMSE (3.6398) and R^2 (3.6398) closest to 1 (least negative R^2 (−0.0076)). CatBoost had the smallest MAE (1.6496)). The highest Directional accuracy of 95.08% was obtained by using Random Forest, which has better performance in determining changes in the GDP growth direction.

All four models showed negative OUT-of-Time R^2 which suggests that none of the models performed better than from this holdout mean benchmark under squared error evaluation. The result sheds light on the challenge of modelling cross-country GDP growth with typical models being based on structural heterogeneity and high economic volatility, decreasing the models' ability to predict growth magnitude correctly.

The significantly lower regression $R^2 = -0.7106$ and RMSE = 4.7425 also supports results from linear regression as linear relationships are not sufficient to capture the non-linear nature of GDP growth dynamics. The ensemble models showed only a minor advantage in performance, indicating however that it is mainly due to the use of nonlinear learning models instead of a given boosting or tree-based one.

Table 1. Comparative performance of machine learning models for out-of-time GDP growth prediction.

Model	R^2	RMSE	MAE	MAPE (%)	Directional accuracy (%)
LightGBM	-0.0076	3.6398	1.7473	132.77	94.62
CatBoost	-0.0093	3.6428	1.6496	118.63	93.54
Random Forest	-0.0171	3.6569	1.7818	133.88	95.08
XGBoost	-0.0521	3.7193	1.8260	132.55	92.78
Linear regression	-0.7106	4.7425	2.9655	148.95	74.65

4.2 Directional accuracy and prediction reliability

The direction of the prediction results are also displayed on Figure 1. The ensemble models outperformed the linear regression model significantly with Random Forest model giving the maximum sign accuracy of 95.08%, followed by LightGBM model with 94.62%, CatBoost model with 93.54% and XGBoost model with 92.78%. A linear regression made a correct growth direction classification of only 74.65%. The directional accuracy is high showing a good fit of the imagery of the broad expansion and contraction patterns of the GDP series, captured by the nonlinear models. However, the interpretation of directional profile need to be taken with caution as sign prediction does not consider the amplitude of the forecasts.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

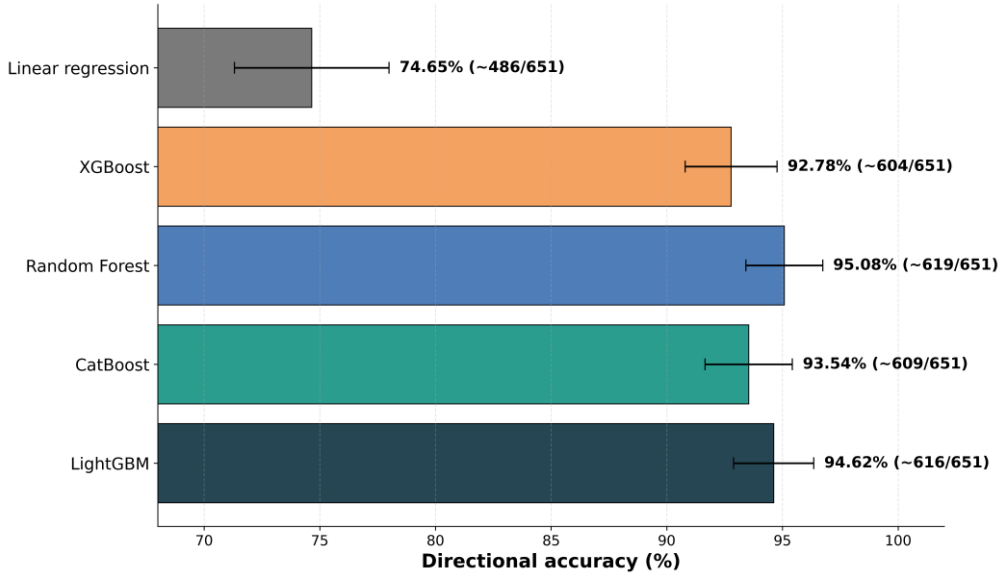

Figure 1. Directional accuracy comparison with descriptive confidence intervals.

A model can accurately signal a period of positive growth, but at the same time seriously misjudge the magnitude of an economic expansion or contraction. Moreover, the uncertainty intervals overlap among ensembles (see Figure 1), which further indicates limited differences between the ensemble models. Thus, the results give evidence for the performance of nonlinear methods of better performance than linear regression for directional classification; however, they can not give strong evidence in favor of the superiority of one ensemble method over the others.

4.3 Relative improvement of nonlinear models over linear regression

Table 2 shows the relative improvements made by each of the ensemble models over linear regression, while the relative improvements are displayed in Figure 2. All non-linear methods resulted in meaningful decreases in errors of prediction. LightGBM gave the biggest reduction (23.25%) in the RMSE while CatBoost gave maximum reduction (44.37%) in the MAE. The highest percentage of reduction in the magnitude of directional error was obtained from Random Forest with a value of 80.59%.

The fact that the results achieved by the nonlinear models are better than those obtained by the linear models in the exploitation of the relationships between the macroeconomic indicators indicates that these models represent a more suitable approach for relationships among variables that are not necessarily orthogonal. The relationships between the macroeconomic indicators and the target GDP growth depend on the state of the growth rates, so they are not required to be orthogonal, as the linear models would have suggested. The differences between ensemble models, however, are not that significant, suggesting it is not an algorithm that matters as much, but moving beyond linear modelling.

Table 2. Relative improvement of nonlinear models compared with linear regression.

Model	Approximate correct signs	95% descriptive interval	RMSE reduction (%)	MAE reduction (%)	Directional-error reduction (%)
LightGBM	~616/651	92.61-96.11%	23.25	41.08	78.78
CatBoost	~609/651	91.39-95.19%	23.19	44.37	74.52

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

Random Forest	~619/651	93.14-96.50%	22.89	39.92	80.59
XGBoost	~604/651	90.53-94.53%	21.58	38.43	71.52
Linear regression	~486/651	71.17-77.84%	Reference	Reference	Reference

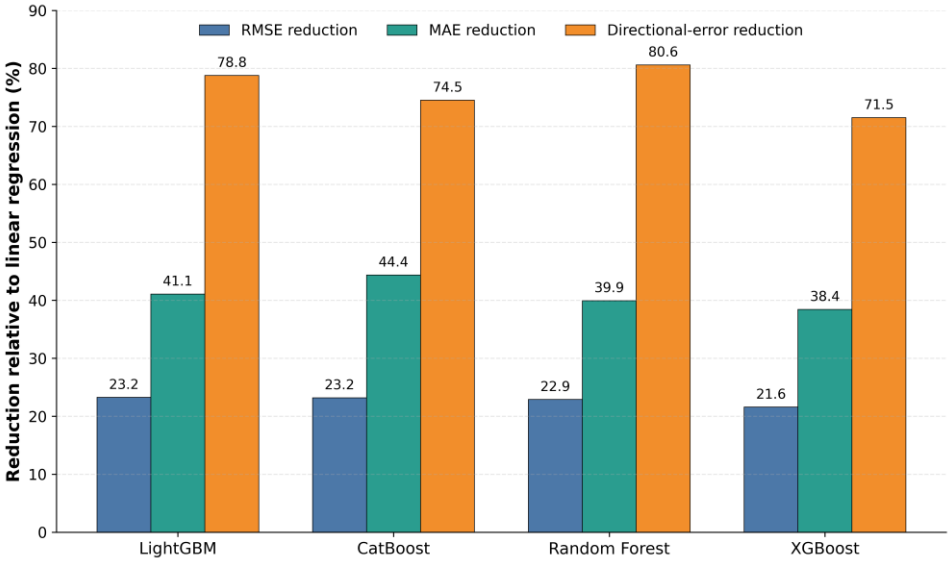

Figure 2. Reduction in RMSE, MAE, and directional error relative to linear regression.

4.4 Prediction behavior and residual characteristics

In Figure 3, the results of the relationship between observed and predicted GDP growth is shown. Model predictions focus on moderate growth values, and with limited spreads from actual results. This suggests that the models reflect general economic features and are not very sensitive to and explanatory of extreme GDP turns. This behaviour is seen in the residual distribution of Figure 4. Prediction errors are clustered near zero, but with long tails, meaning that there are often large errors related to very unusual economic events. For most of these errors, they are extreme errors, and these extreme errors are part of why we have these large negative R^2 values in Table 1: squared-error measures do penalize large forecasting errors very heavily. Overall, conclusions suggest an improvement in the average prediction of the ensemble models while also noting that they fall short in the prediction of rare macroeconomic shocks and even country-specific disruptions.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

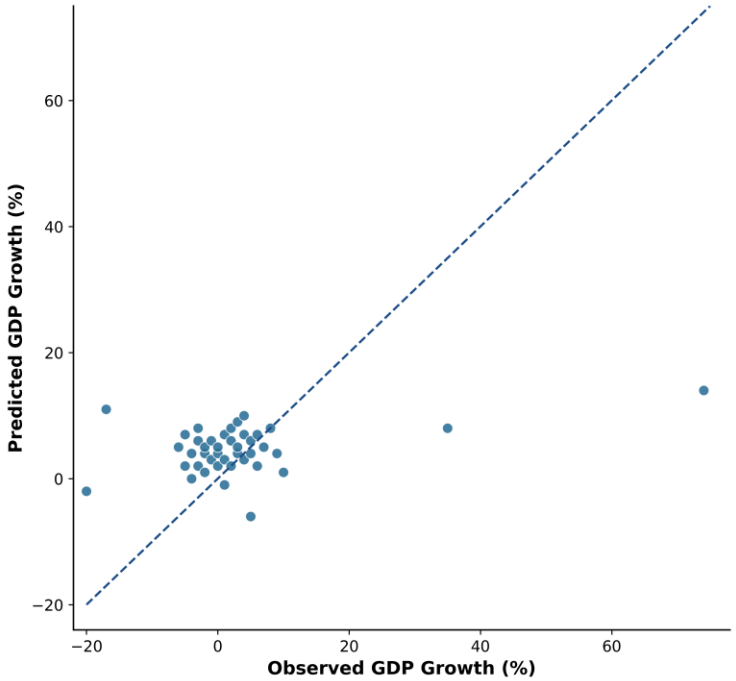

Figure 3. Observed versus predicted GDP growth values.

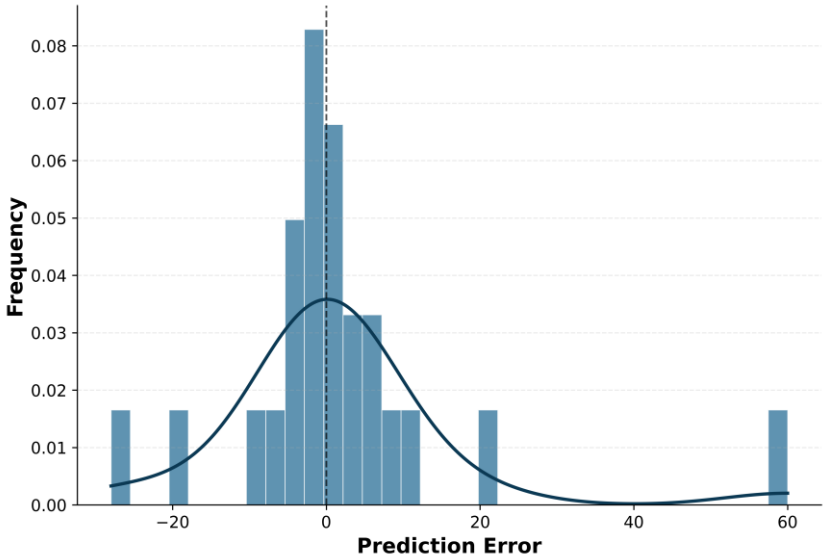

Figure 4. Distribution of prediction errors.

4.5 Explainable AI analysis of influential predictors

Figure 5 and Figure 6 show result of analyzing Feature importance for LightGBM. In fact, lagged GDP growth was shown to be the most important indicator, followed by current account balance and inflation related indicators in the split-based importance analysis. It indicates that prior performance of the economy and the external environment are good predictors of GDP growth. The ranking by SHAP, however, is different, showing that the sensible people with the highest average contribution are factors related to GDP per capita and information of the year. These two methods highlight the differences between

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

methodologies and the importance should not be used to rank economics as this is not a final rule. Based on the results, forecasting of GDP growth is based on both historical persistence and level of development together with macroeconomic conditions, and no single factor is more significant. These explanations refer to model behavior and are not causal explanations.

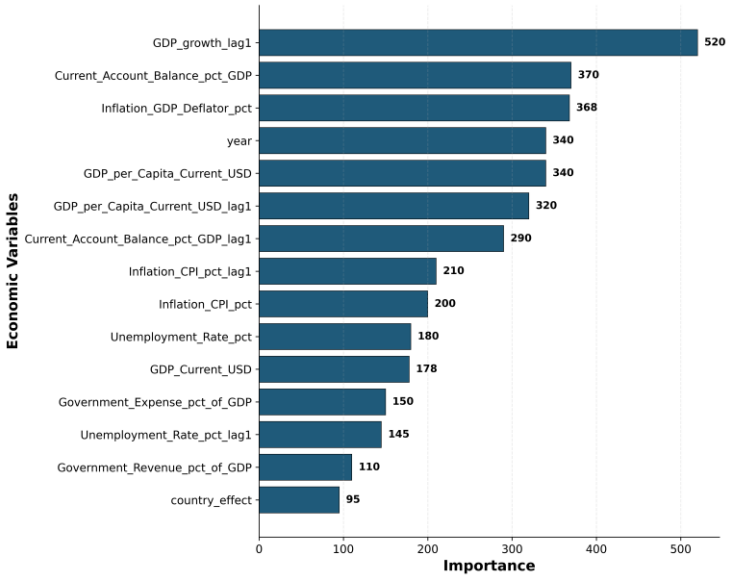

Figure 5. LightGBM split-based feature importance.

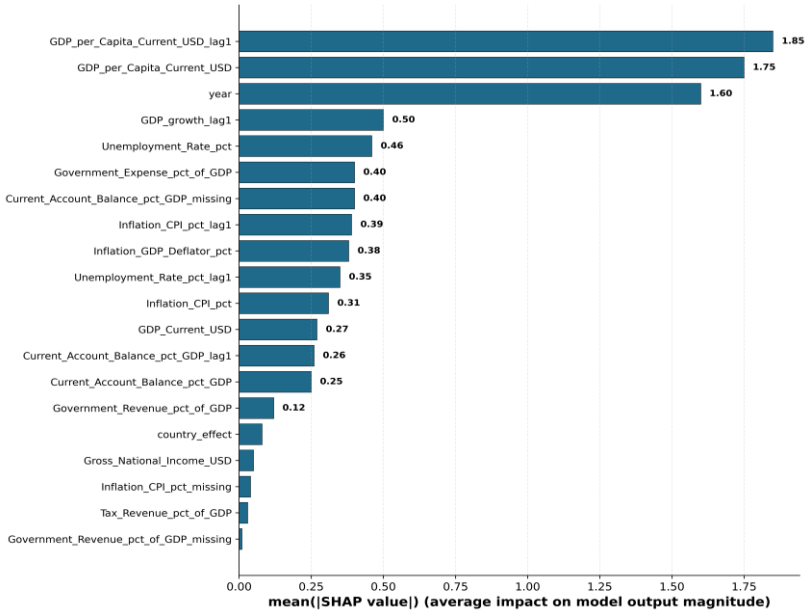

Figure 6. Mean absolute SHAP feature importance.

4.6 Nonlinear response patterns of key predictors

The partial dependence plots in Figures 7 and 8 show a non-linear relationship between some of the predictors and the predicted growth in GDP. The response curve for lagged GDP growth shows threshold-like changes, indicating the model does not imply a constant marginal effect of the GDP growth. The same holds true for the current-account balance relationship where predicted GDP growth is found to have a nonlinear relationship. The patterns seem to

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

reflect the capacity of tree-based models to capture interactions and discontinuities in the economic relationships. But these curves are modelling behavior, not structural effects. The thresholds that were observed were not necessarily economic turning points, but were related to the spread of data, country difference or tree partitioning.

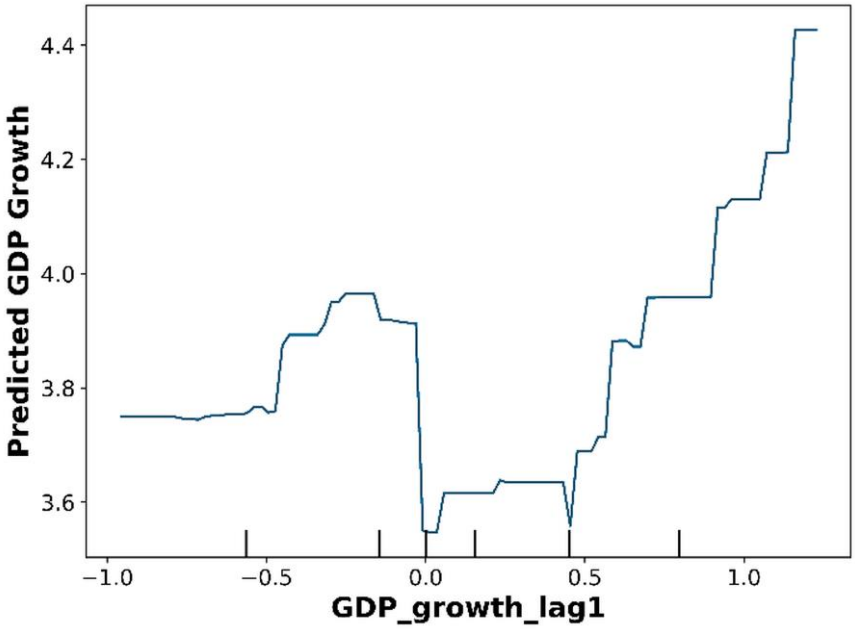

Figure 7. Partial dependence of predicted GDP growth on lagged GDP growth.

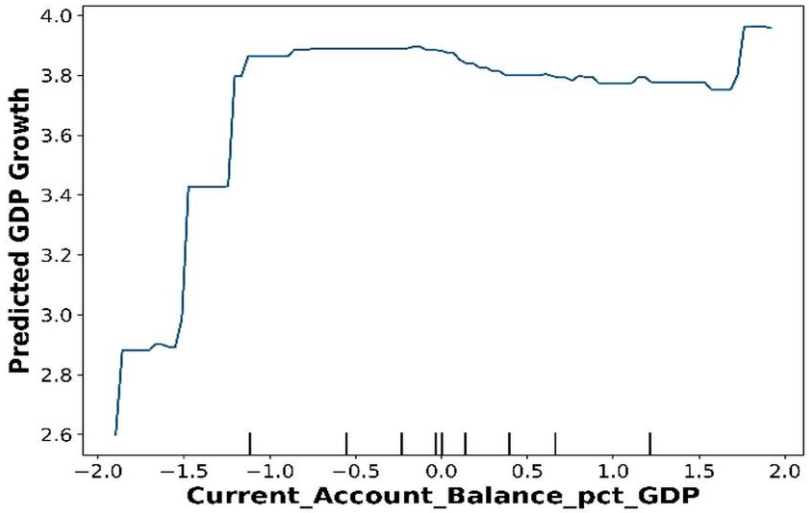

Figure 8. Partial dependence of predicted GDP growth on current-account balance.

4.7 Cross-country variation in prediction errors

Figure 9 shows the distribution of the respective prediction errors obtained with respect to country. However, there are significant differences between the economic prediction errors of all the economies, with Macao SAR having the greatest average error of all the countries shown. The prediction uncertainty is also high for a couple of smaller and less-stable economies. This heterogeneity identifies a major weakness of global GDP modelling: the same modelling framework can work differently in economies that have different structural properties, have more or less data and are differently exposed to shocks.

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

Country-specific calibration, hierarchical modelling and uncertainty-aware forecasting method shall be explored in future studies.

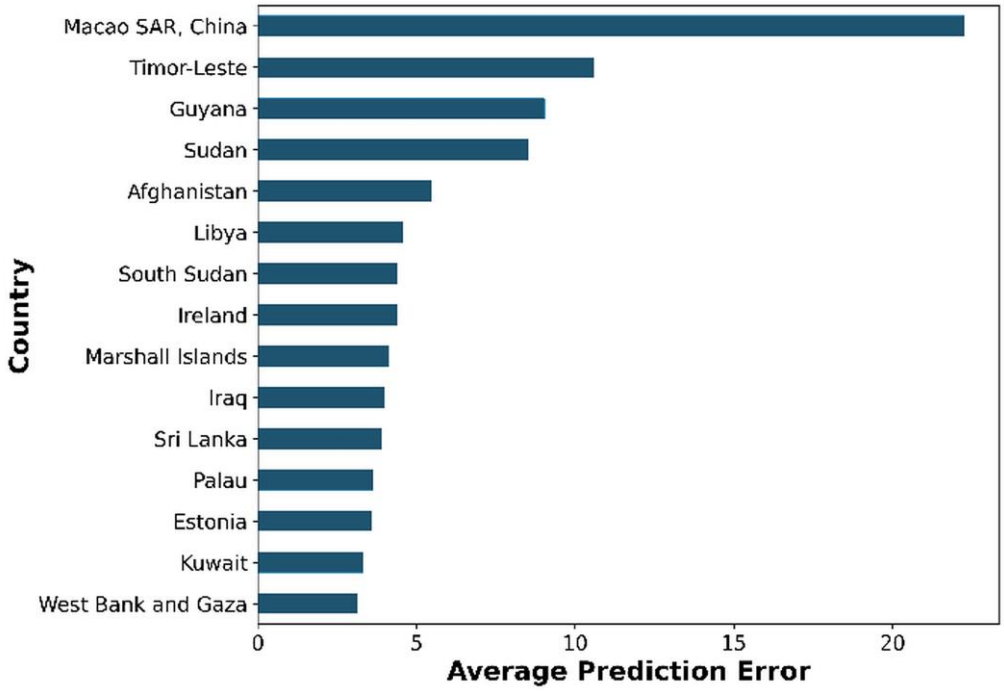

Figure 9. Countries with highest average prediction errors.

5.0 Discussion

5.1 Nonlinear machine learning improves GDP growth modelling but does not guarantee superior forecasting

The results show that nonlinear models of the ensemble of models are able to offer significant improvement in predicting cross-country GDP growth when compared to linear regression. The results show that LightGBM, CatBoost, Random Forest, and XGBoost models have lower performance on the RMSE, MAE, and directional error scores, suggesting that these models better capture complex interaction between the various macroeconomic variables. These findings are found to be consistent with the situations of nonlinearity, country heterogeneity, and interaction effects which are hard to capture with the conventional linear specifications of economic growth dynamics.

The negative R^2 values for all models however, suggest that better relative performance does not lead to good out-of-time forecasting skills. The LightGBM model, which has achieved great success using squared error metric, still lagged slightly behind the holdout mean metric. This is a major difference between comparing models and forecasting them absolutely. Yet, the ensemble models performed better than the linear regression has been predicting growth in GDP but it is still uncertain. Thus, rather than replacing the traditional frameworks of forecasting, machine learning is helping bring the nonlinear patterns out of large macroeconomic datasets.

5.2 Directional accuracy and numerical forecasting reveal different aspects of model performance

One of the findings of this study is that while there is a significant difference between the accuracy of numerical prediction and the accuracy of directional accuracy. Random Forest achieved the best prediction of 'sign? (95.08%) and LightGBM attained the best RMSE score. The difference represents an example of the fact that a valuation's absolute level does not necessarily follow from its direction of movement.

This behaviour is understandable from the figure of prediction compression that is

JOURNAL FOR SOCIAL SCIENCE STUDIES
An International Peer-Reviewed Open Access Journal
ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

observed in Figure 3. The models are expected to be more accurate around average economic conditions - setting the correct sign is more important during normal economic expansion periods, but is less useful to explain extreme events. Likewise the residuals in Figure 4 indicate that growth variation, especially extreme growth deviations, are hard to forecast due to the clustering of large errors in the tails. These results highlight the need to measure the performance of GDP forecasting models by several indicators. Directional accuracy might be considered for any economic indicator where having a rough idea of a trend is sufficient; however, RMSE and MAE are still necessary metrics for applications that need close quantitative forecasts.

5.3 Explainable AI reveals predictive signals rather than causal economic mechanisms

The explainability analysis gives them an insight into how the LightGBM model makes its predictions but does not provide any evidence of causal economic relationships. Lagged GDP growth, currentaccount indicators, inflation measures and income related variables were found to be important predictors in the feature importance analysis. The results confirm that historical economic conditions/structural development characteristics carry information that can be useful for forecasting subsequent structural growth.

Importance and SHAP values, however, show that interpretation is highly dependent on the choice of explanation method used, which is why it is important to compare split-based importance with SHAP scores. With lagged GDP growth, the dominant importance factors are the tree-based ones, whereas the importance factors for the GDP per capita variables and year-related effects are highlighted by SHAP. This difference may be due to the correlations between the so called macroeconomic indicators, and the various ways in which the importance was being quantified. This means that the use of explainable AI techniques as diagnostics for understanding model behavior should be taken as a snapshot of model behavior, and not as proof of economic drivers. This study suggested an economic interpretation, which would need further identification techniques, causal models, and robustness analyses.

5.4 Implications for global GDP monitoring and future research

Not only is the analytical framework of nonlinear ensemble models valuable in studying the global GDP growth rate, but the results also indicate that it is the use of multiple interacting indicators within the datasets, combined with the diverse economies within the essentials, that renders the use of such a framework worthwhile. The improvement over linear regression over repeated results is sufficient for testing the application of flexible machine learning tools as exploratory tools for monitoring and comparative power parameters in macroeconomics.

However, there are a number of enhancements needed before such models can be used in a predictive forecasting system. Future studies should include real time vintages of data to mitigate information availability concerns, perform model performance comparisons to naïve forecasts, such as historical persistence and mean forecasts, and assess models on several rolling forecast origins. Furthermore, the results could be enhanced by conducting country-specific uncertainty analysis and by employing hierarchical modelling for some economies that do not have large amounts of data and/or have a high degree of volatility.

In all, this research shows that relationship of the GDP between countries is nonlinear, which can be captured by the machine learning ensembles, but only when it is properly validated and interpreted. Overall, the evidence does not support the use of machine learning as an alternative to more traditional macroeconomic forecasting methods, but rather as a complementary method.

5.5 Limitations

- The evaluation uses one chronological holdout, so model rankings may be period-specific.

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

- Same-year predictors and undocumented release dates prevent a verified real-time forecasting interpretation.
- The preprocessing fit scope cannot be independently checked from the supplied manuscript and images.
- No naive mean, persistence or majority-sign results are reported alongside the fitted models.
- Observation-level predictions are unavailable, preventing paired accuracy tests, clustered bootstrap intervals, calibration analysis and error decomposition by year.
- MAPE is unstable near zero and for negative growth, reducing its usefulness for annual GDP changes.
- Country and year integer encodings are predictive proxies, not equivalent to econometric fixed effects.
- Feature-attribution results are sensitive to correlated predictors and the selected explanation method.
- The inclusion of 2025 requires a clear data-extraction date and distinction between observed values, provisional estimates and missing entries.

6.0 Conclusion

This study demonstrates that nonlinear tree-based ensembles provide superior predictive performance compared with linear regression for modelling global GDP growth dynamics. The ensemble models captured complex nonlinear relationships within macroeconomic indicators, achieving substantial improvements in RMSE, MAE, and directional error reduction. However, the negative out-of-time R^2 values indicate that improved model fit relative to linear regression does not necessarily imply superior forecasting beyond simple benchmarks. The findings highlight the importance of multi-dimensional evaluation, as directional accuracy and numerical prediction performance capture different aspects of forecasting ability. Explainable AI analysis further reveals that historical growth, income-related indicators, and macroeconomic conditions contribute strongly to model predictions; however, these relationships represent predictive associations rather than causal effects. Overall, machine learning ensembles provide a valuable framework for extracting nonlinear patterns from cross-country macroeconomic data, but their application requires robust validation, appropriate benchmarks, and careful interpretation. Future research should focus on real-time forecasting designs, uncertainty estimation, and country-specific modelling approaches to enhance the reliability of GDP growth prediction systems.

Data availability statement

The source indicators are publicly available through the World Bank World Development Indicators database, including GDP growth (annual %) at https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG. The processed analytical file identified in this study as world_bank_data_2025.csv, the exact indicator-extraction record, code and observation-level predictions are not currently archived in a public repository. Requests for these materials should be directed to the corresponding author.

Code availability statement

The executable preprocessing and model-training code was not included with the manuscript materials reviewed for this revision. A reproducible release should include package versions, random seeds, temporal split logic, fitted preprocessing scope and scripts used to produce every table and figure.

Funding

The authors received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest

The authors declare that they have no competing interests.

Ethics statement

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

Ethical approval was not required because the study uses publicly available, aggregate country-level economic indicators and involves no human participants or personal data.

REFERENCES

Smets, F. and Wouters, R. (2007) 'Shocks and frictions in US business cycles: A Bayesian DSGE approach', American Economic Review, 97(3), pp. 586-606. https://doi.org/10.1257/aer.97.3.586.

Edge, R.M. and Gurkaynak, R.S. (2010) 'How useful are estimated DSGE model forecasts for central bankers?', Brookings Papers on Economic Activity, 2010(2), pp. 209-259.

Hendry, D.F. and Clements, M.P. (2004) 'Pooling of forecasts', The Econometrics Journal, 7(1), pp. 1-31.

Giannone, D., Reichlin, L. and Small, D. (2008) 'Nowcasting: The real-time informational content of macroeconomic data', Journal of Monetary Economics, 55(4), pp. 665-676. https://doi.org/10.1016/j.jmoneco.2008.05.010.

Koop, G. and Korobilis, D. (2013) 'Large time-varying parameter VARs', Journal of Econometrics, 177(2), pp. 185-198. https://doi.org/10.1016/j.jeconom.2013.04.007.

Breiman, L. (2001) 'Random forests', Machine Learning, 45, pp. 5-32. https://doi.org/10.1023/A:1010933404324.

Biau, G. and Scornet, E. (2016) 'A random forest guided tour', TEST, 25(2), pp. 197-227. https://doi.org/10.1007/s11749-016-0481-7.

Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', The Annals of Statistics, 29(5), pp. 1189-1232. https://doi.org/10.1214/aos/1013203451.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 785-794. https://doi.org/10.1145/2939672.2939785.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) 'LightGBM: A highly efficient gradient boosting decision tree', Advances in Neural Information Processing Systems, 30, pp. 3146-3154.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) 'CatBoost: Unbiased boosting with categorical features', Advances in Neural Information Processing Systems, 31, pp. 6638-6648.

Tashman, L.J. (2000) 'Out-of-sample tests of forecasting accuracy: An analysis and review', International Journal of Forecasting, 16(4), pp. 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0.

Bergmeir, C., Hyndman, R.J. and Koo, B. (2018) 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', Computational Statistics & Data Analysis, 120, pp. 70-83. https://doi.org/10.1016/j.csda.2017.11.003.

Chakraborty, C. and Joseph, A. (2017) Machine learning at central banks. Bank of England Staff Working Paper No. 674. London: Bank of England.

Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence, 1(5), pp. 206-215. https://doi.org/10.1038/s42256-019-0048-x.

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

Apley, D.W. and Zhu, J. (2020) 'Visualizing the effects of predictor variables in black box supervised learning models', Journal of the Royal Statistical Society: Series B, 82(4), pp. 1059-1086. https://doi.org/10.1111/rssb.12377.

Zhao, Q. and Hastie, T. (2021) 'Causal interpretations of black-box models', Journal of Business & Economic Statistics, 39(1), pp. 272-281. https://doi.org/10.1080/07350015.2019.1624293.

Varian, H.R. (2014) 'Big data: New tricks for econometrics', Journal of Economic Perspectives, 28(2), pp. 3-28. https://doi.org/10.1257/jep.28.2.3.

Mullainathan, S. and Spiess, J. (2017) 'Machine learning: An applied econometric approach', Journal of Economic Perspectives, 31(2), pp. 87-106. https://doi.org/10.1257/jep.31.2.87.

Athey, S. and Imbens, G.W. (2019) 'Machine learning methods that economists should know about', Annual Review of Economics, 11, pp. 685-725. https://doi.org/10.1146/annurev-economics-080217-053433.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018) 'Statistical and machine learning forecasting methods: Concerns and ways forward', PLOS ONE, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889.

Giannone, D., Lenza, M. and Primiceri, G.E. (2021) 'Economic predictions with big data: The illusion of sparsity', Econometrica, 89(5), pp. 2409-2437. https://doi.org/10.3982/ECTA17897.

Goulet Coulombe, P., Leroux, M., Stevanovic, D. and Surprenant, S. (2022) 'How is machine learning useful for macroeconomic forecasting?', Journal of Applied Econometrics, 37(5), pp. 920-964. https://doi.org/10.1002/jae.2910.

Goulet Coulombe, P. (2020) 'The macroeconomy as a random forest', Quantitative Economics, 11(4), pp. 1279-1317. https://doi.org/10.3982/QE1342.

Medeiros, M.C., Vasconcelos, G.F.R., Veiga, A. and Zilberman, E. (2021) 'Forecasting inflation in a data-rich environment: The benefits of machine learning methods', Journal of Business & Economic Statistics, 39(1), pp. 98-119.

Baltagi, B.H. (2008) Econometric analysis of panel data. 4th edn. Chichester: John Wiley & Sons.

Diebold, F.X. and Mariano, R.S. (1995) 'Comparing predictive accuracy', Journal of Business & Economic Statistics, 13(3), pp. 253-263. https://doi.org/10.1080/07350015.1995.10524599.

Pesaran, M.H. and Timmermann, A. (1992) 'A simple nonparametric test of predictive performance', Journal of Business & Economic Statistics, 10(4), pp. 461-465. https://doi.org/10.1080/07350015.1992.10509922.

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems, 30, pp. 4765-4774.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020) 'From local explanations to global understanding with explainable AI for trees', Nature Machine Intelligence, 2, pp. 56-67. https://doi.org/10.1038/s42256-019-0138-9.

World Bank (2026) GDP growth (annual %). World Development Indicators. Available at: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG (Accessed: 16 September 2026).

ISSN (Online): 3006-2071 | ISSN (Print): 3006-2063
Volume 4 • Issue 5 • September 2026

Hyndman, R.J. and Koehler, A.B. (2006) 'Another look at measures of forecast accuracy', International Journal of Forecasting, 22(4), pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.