UDC: 330.35:004.85
DOI: https://doi.org/10.30546/09090.2026.001.20.2052

STACKED ENSEMBLE LEARNING FOR ECONOMIC GROWTH FORECASTING: A CROSS-COUNTRY PANEL DATA ANALYSIS

Parviz JABRAYILLI
Azerbaijan State Oil and Industry University
Baku Azerbaijan

ARTICLE INFO

Article history:
Received: 2025-11-21
Received in revised form:2025-12-11
Accepted:2026-01-10
Available online:2026-06-23

Keywords:
Economic Growth Forecasting;
Ensemble learning;
Machine learning;
Stacking;
Panel data;

2010 Mathematics Subject Classifications: 68T05, 62M20, 62P20, 91B62

ABSTRACT

Economic growth forecasting remains a fundamental challenge in macroeconomic analysis, with traditional econometric models often struggling to capture complex non-linear relationships and structural heterogeneity across countries. This study develops and evaluates stacked ensemble machine learning models for predicting real GDP growth rates across a diverse panel of 85 countries spanning developed, emerging, and developing economies over the period 1990–2022. Multiple base learners including Random Forest, XGBoost, LightGBM, CatBoost, Gradient Boosting, and Support Vector Regression are combined through stacking architecture with Ridge regression as the meta-learner. Out-of-sample forecasting results demonstrate that the stacked ensemble achieves superior predictive accuracy with RMSE of 1.42%, MAE of 1.08%, and R^2 of 0.877, reducing forecast error by 15.7% relative to the best single model and 41.7% relative to ARDL specifications. SHAP value analysis reveals that capital formation, human capital, and trade openness emerge as the most influential predictors, with substantial non-linear effects and cross-country heterogeneity. These findings suggest that ensemble learning frameworks offer significant advantages for economic growth forecasting by effectively capturing complex interactions and structural differences across economies.

1. INTRODUCTION

Accurate forecasting of economic growth represents a cornerstone of macroeconomic analysis, informing policy decisions, investment strategies, and international development planning [1,2]. The complexity inherent in economic growth processes, characterized by non-linear relationships, structural breaks, cross-country heterogeneity, and multidimensional interactions among determinants, poses substantial challenges for traditional econometric modeling frameworks. While classical approaches such as autoregressive distributed lag (ARDL) models, vector autoregression (VAR), and dynamic panel estimators have served as workhorses in growth empirics, their parametric assumptions and limited capacity to capture complex non-linearities increasingly constrain predictive performance in contemporary data-rich environments [4-6].

The proliferation of machine learning methodologies in economic forecasting has opened new avenues for addressing these limitations [8]. Machine learning algorithms excel at discovering

complex patterns in high-dimensional data without imposing rigid functional form assumptions, making them particularly suited for growth prediction where theoretical guidance on functional relationships remains incomplete. Recent applications of Random Forests, Gradient Boosting, and Neural Networks to macroeconomic forecasting have demonstrated promising improvements over traditional methods [9-11,14,15]. However, single machine learning models exhibit inherent limitations, with each algorithm possessing distinct strengths and weaknesses in capturing different aspects of the data-generating process. Random Forests excel at handling non-linearities but may underperform in capturing smooth trends, while Gradient Boosting methods achieve high accuracy but risk overfitting.

Ensemble learning methods, which systematically combine predictions from multiple base models, have emerged as a powerful paradigm for enhancing predictive accuracy by leveraging the complementary strengths of diverse algorithms [13,21]. In contrast to single models, ensembles reduce prediction variance, mitigate algorithm-specific biases, and improve robustness across different data regimes. Stacking, an advanced ensemble technique that employs a meta-learner to optimally weight base model predictions, has shown particular promise in domains requiring high predictive accuracy [12,22]. Despite extensive application in weather forecasting, healthcare analytics, and financial prediction, the systematic evaluation of stacked ensemble learning for cross-country economic growth forecasting remains notably absent from the literature.

Existing research on machine learning applications in growth economics exhibits several limitations that this study addresses. First, most studies focus on single-country contexts or small panels, limiting generalizability across diverse economic structures [17]. Second, comparative analyses typically evaluate only a narrow subset of algorithms, failing to explore the potential of combining heterogeneous models through sophisticated ensemble architectures [14]. Third, previous work often emphasizes in-sample fit rather than rigorous out-of-sample forecasting performance, which represents the true test of predictive models [15]. Fourth, limited attention has been devoted to interpreting model predictions and understanding the economic mechanisms driving forecast accuracy, with recent advances in explainable AI remaining underutilized in macroeconomic applications [16].

This study aims to develop and evaluate stacked ensemble machine learning frameworks for economic growth forecasting across a heterogeneous panel of 85 countries spanning different income levels, geographic regions, and development stages over the period 1990–2022. The specific research objectives are threefold. First, to construct and compare the out-of-sample predictive performance of multiple single machine learning models against traditional econometric benchmarks. Second, to develop stacked ensemble architectures that optimally combine base learners and evaluate their forecasting superiority through rigorous statistical tests. Third, to employ SHAP (SHapley Additive exPlanations) value analysis to interpret ensemble predictions and identify the most influential growth determinants across countries and time periods.

The contributions of this research are both methodological and substantive. Methodologically, this study represents the first comprehensive application of stacked ensemble learning to cross-country economic growth prediction, demonstrating the substantial gains achievable through optimal model combination. The analysis incorporates a diverse set of eight base learners spanning tree-based methods, boosting algorithms, and kernel methods, combined through stacking with systematic hyperparameter optimization via Bayesian methods. Substantively, the research provides new evidence on the determinants of economic growth, revealing complex non-linear relationships and cross-country heterogeneity that traditional linear models cannot

capture. The SHAP value analysis offers novel insights into how growth determinants operate differently across development stages and economic structures. Practically, the superior forecasting accuracy demonstrated by ensemble methods has direct implications for policy planning, investment decisions, and international development strategy formulation.

The remainder of this paper is organized as follows. Section 2 reviews the literature on economic growth modeling, machine learning applications in macroeconomic forecasting, and ensemble learning methods. Section 3 describes the data sources, variable construction, and panel composition. Section 4 presents the methodology, detailing base learners, stacking architecture, hyperparameter optimization, and evaluation metrics. Section 5 reports empirical results including performance comparisons, feature importance analysis, and robustness checks. Section 6 discusses the implications of findings, compares results with previous literature, and acknowledges limitations. Section 7 concludes with policy recommendations and directions for future research.

2. LITERATURE REVIEW

2.1. Traditional Econometric Approaches to Growth Modeling

Economic growth modeling has evolved substantially since the seminal contributions of Solow and Swan, who formalized the neoclassical growth framework emphasizing capital accumulation, labor force growth, and exogenous technological progress [1]. Subsequent theoretical advances introduced endogenous growth mechanisms through human capital, knowledge spillovers, and institutional quality [2,3]. Empirically, the growth literature has predominantly relied on cross-sectional regressions, panel data estimators, and time series models to identify growth determinants and forecast future trajectories.

The autoregressive distributed lag (ARDL) framework has become a standard tool for growth forecasting, particularly when integrating variables exhibit different orders of integration [4]. ARDL models accommodate short-run dynamics and long-run equilibrium relationships while permitting flexible lag structures. However, these models impose strong parametric assumptions, assume linear relationships, and struggle with high-dimensional predictor spaces. Vector autoregression (VAR) methods provide multivariate alternatives but face dimensionality curses as the number of variables increases, while structural interpretations require contentious identifying restrictions.

Dynamic panel estimators including difference GMM and system GMM address endogeneity concerns in growth regressions by exploiting internal instruments [5,6]. These methods have generated extensive insights into convergence dynamics and policy effects. Nevertheless, they remain linear in parameters, assume homogeneous slope coefficients across countries, and provide limited forecasting accuracy when non-linearities and structural heterogeneity characterize the data-generating process. Recent meta-analyses document substantial heterogeneity in estimated growth effects across studies, suggesting that traditional models may inadequately capture the complex, context-dependent nature of growth processes [7,25].

2.2. Machine Learning Applications in Macroeconomic Forecasting

The integration of machine learning methodologies into macroeconomic forecasting has accelerated over the past decade, driven by increasing data availability and computational advances [8]. Recent studies demonstrate that machine learning models substantially outperform traditional methods for nowcasting GDP growth using high-frequency indicators, achieving 30–40% reductions in forecast errors [15].

Random Forests have found extensive applications in macroeconomic prediction [9]. Applications to predict US GDP growth using hundreds of predictors demonstrate superior performance over factor models and LASSO regressions [14]. The algorithm's ability to handle non-linearities, interactions, and irrelevant variables without explicit specification makes it particularly attractive for growth forecasting where theoretical guidance on functional forms remains incomplete.

Gradient Boosting methods represent another prominent machine learning approach in economic applications [10,11]. These algorithms iteratively construct ensembles of weak learners, with each successive model focusing on observations poorly predicted by previous iterations. Applications to forecasting inflation and GDP growth demonstrate substantial accuracy gains over AR and ARIMA benchmarks [17,18]. The success of boosting methods stems from their capacity to capture complex non-linear relationships while maintaining regularization through shrinkage and tree depth constraints.

Neural Networks and deep learning architectures have also attracted attention for macroeconomic forecasting [19]. However, neural networks require substantial data for training, face challenges in hyperparameter selection, and provide limited interpretability, constraining their adoption for policy-relevant growth forecasting where understanding drivers is crucial.

While these studies demonstrate machine learning's potential for improving forecast accuracy, they predominantly focus on single algorithms applied to individual countries or small panels. Comparative evaluations across diverse algorithms remain limited, and most research emphasizes in-sample fit rather than rigorous out-of-sample forecasting performance.

2.3. Ensemble Learning Methods and Their Applications

Ensemble learning, which combines predictions from multiple base models to achieve superior accuracy, has emerged as a dominant paradigm in machine learning competitions and practical applications [21]. The fundamental principle underlying ensembles is that diverse models make different errors, and appropriate combination reduces overall prediction variance while maintaining or reducing bias. Theoretical foundations for ensemble learning were established for bagging and boosting, demonstrating conditions under which ensembles provably outperform individual models [11,13].

Simple ensemble methods include voting and weighted averaging. Bagging (Bootstrap Aggregating), employed in Random Forests, reduces variance by training models on bootstrap samples of the data. Boosting methods sequentially train models to correct previous errors, effectively reducing both bias and variance.

Stacking represents a more sophisticated ensemble approach that uses a meta-learner to combine base model predictions [12]. In stacking, diverse base learners (Level-0 models) are trained on the data, and their predictions serve as inputs to a meta-learner (Level-1 model) that learns optimal combination weights. This architecture enables the meta-learner to exploit situations where different base models excel, assigning higher weights to more reliable predictions for specific input regions [22].

Empirical applications of ensemble learning span numerous domains. In weather forecasting, ensemble prediction systems combining multiple atmospheric models have become standard practice, substantially improving forecast accuracy and uncertainty quantification. In healthcare, ensemble methods achieve state-of-the-art performance for disease prediction and medical image analysis [21]. Financial applications include credit risk assessment, stock market prediction, and fraud detection, consistently demonstrating that ensembles outperform single models.

In economics and forecasting, ensemble applications remain relatively nascent but growing. Early work pioneered the use of forecast combination methods for macroeconomic prediction, demonstrating that simple averages of individual forecasts often outperform sophisticated models [20]. Recent research has begun exploring machine learning ensembles for macroeconomic forecasting [14,15]. Despite these advances, systematic evaluation of stacking architectures for cross-country economic growth prediction remains absent, representing a significant gap in the literature.

2.4. Research Gap

The systematic literature review reveals several critical gaps that this study addresses. First, while machine learning applications in macroeconomic forecasting have proliferated, most research focuses on single-country contexts or small panels of developed economies. Second, comparative analyses typically evaluate a limited subset of algorithms. Third, existing studies predominantly emphasize in-sample performance metrics. Fourth, interpretability of machine learning predictions receives insufficient attention. This study fills these gaps by developing and evaluating stacked ensemble learning frameworks for economic growth forecasting across a heterogeneous panel of 85 countries, incorporating eight diverse base learners combined through stacking with systematic hyperparameter optimization and SHAP value analysis for interpretation.

3. DATA AND VARIABLES

3.1. Data Sources and Country Selection

The empirical analysis employs a balanced panel dataset comprising 85 countries observed annually over the period 1990–2022, yielding 2,805 country-year observations. The panel composition was designed to ensure representation across diverse economic structures, development stages, geographic regions, and institutional contexts. Country selection followed three criteria: data availability, economic diversity (28 high-income OECD countries, 32 upper-middle-income economies, and 25 lower-middle and low-income countries), and geographic representation spanning six continents.

Macroeconomic data were sourced from multiple authoritative databases. Real GDP, capital stock, employment, and demographic variables were obtained from the Penn World Table (PWT) version 10.0 [23]. Trade openness, foreign direct investment, inflation, and monetary aggregates were extracted from the World Bank's World Development Indicators (WDI) database. Institutional quality indicators were sourced from the Worldwide Governance Indicators (WGI) project [24]. Human capital indices were obtained from the Human Development Index (HDI) database.

3.2. *Variable Construction*

The dependent variable is the annual real GDP growth rate, calculated as:

$$g_{it} = \ln(GDP_{it}) - \ln(GDP_{it-1})$$

where i indexes countries and t indexes years. Real GDP growth exhibits substantial cross-country and temporal variation, ranging from -15.2% to 22.4%, with mean 3.8% and standard deviation 4.2% across the sample.

Explanatory variables encompass 18 macroeconomic and institutional determinants selected based on theoretical foundations in growth economics. Physical capital is measured as capital stock per worker in constant 2017 purchasing power parity dollars. Investment rate represents gross fixed capital formation as percentage of GDP. Labor force captures total employment in

thousands. Human capital index incorporates average years of schooling and returns to education. Trade openness is the sum of exports and imports as percentage of GDP. Foreign direct investment represents net FDI inflows as percentage of GDP. Inflation rate measures annual percentage change in consumer price index. Additional variables include government expenditure, population growth, urbanization rate, financial development, terms of trade index, and four governance indicators (effectiveness, regulatory quality, rule of law, corruption control).

3.3. Data Treatment

Data preprocessing was conducted using a leakage-safe pipeline designed to preserve temporal structure and ensure reproducibility. Missing values (approximately 0.8% of observations) were imputed using linear interpolation within each country time series prior to model partitioning to maintain chronological consistency. Outliers were examined using the interquartile range criterion; extreme values were retained because they corresponded to genuine macroeconomic events rather than measurement errors.

Stationarity was evaluated using panel unit root diagnostics. Variables identified as non-stationary were transformed through first differencing applied separately within each country series to preserve panel integrity and avoid cross-sectional information contamination.

Feature scaling was performed using z-score standardization. Importantly, scaling parameters (mean and standard deviation) were estimated **exclusively from the training period (1990–2012)**. These fixed parameters were then applied unchanged to the validation (2013–2017) and test sets (2018–2022). This protocol prevents information leakage from future observations into model training.

The dependent variable — annual real GDP growth — was **not standardized**. Forecast accuracy metrics (RMSE, MAE, and MAPE) were computed directly in percentage-point units to preserve economic interpretability, eliminating the need for inverse transformations.

The dataset was partitioned strictly by time into training, validation, and test sets. All preprocessing transformations were fitted using training data only and applied consistently across subsequent periods, ensuring that reported forecasting performance reflects genuine out-of-sample predictive capability.

Table 1. *Descriptive Statistics of Key Variables*

Variable	Mean	Std. Dev.	Min	Max	Between SD	Within SD
GDP Growth (%)	3.84	4.23	-15.21	22.43	3.89	2.67
Capital per Worker (Constant 2017 PPP, 000 USD)	127.4	98.3	8.2	456.7	95.1	18.6
Investment Rate (% GDP)	23.4	6.8	10.2	48.6	6.1	3.4
Human Capital Index	2.68	0.52	1.24	3.87	0.49	0.14
Trade Openness (% GDP)	82.6	58.4	18.4	442.8	56.7	12.8
FDI (% GDP)	3.82	5.73	-12.4	45.7	3.21	4.89
Inflation (%)	8.41	16.3	-9.8	135.2	8.7	14.1
Governance Effectiveness	0.02	0.98	-2.48	2.37	0.95	0.21

Note: All growth-related variables and error metrics are expressed in percentage points unless otherwise stated.

4. Methodology

4.1. Overview of Machine Learning Framework

The methodological framework employs a two-stage stacked ensemble architecture combining diverse base learners through a meta-learner. This approach leverages the complementary strengths of multiple algorithms while mitigating individual model weaknesses. The modeling pipeline consists of four sequential steps: base learner training and optimization, cross-validated prediction generation, meta-learner training, and final ensemble prediction on test data.

4.2. *Base Learners*

The ensemble incorporates eight diverse base learners spanning tree-based methods, gradient boosting algorithms, kernel methods, and linear models.

Random Forest (RF) constructs an ensemble of decision trees through bootstrap aggregating combined with random feature selection. Final predictions aggregate individual tree predictions through averaging.

where B is the number of trees. Random Forests excel at capturing non-linear relationships and interactions without explicit specification. Key hyperparameters optimized include number of trees, maximum tree depth, minimum samples per leaf, and number of features per split.

Extreme Gradient Boosting (XGBoost) implements gradient boosting through sequential construction of decision trees. The objective function combines prediction error and regularization.

where l is a differentiable loss function and Ω penalizes tree complexity. XGBoost employs Newton boosting with second-order Taylor approximation, achieving faster convergence and better regularization than traditional gradient boosting.

Light Gradient Boosting Machine (LightGBM) introduces histogram-based gradient boosting with leaf-wise tree growth strategy, achieving substantial computational efficiency improvements. The algorithm bins continuous features into discrete bins, reducing memory consumption and accelerating split finding.

CatBoost implements gradient boosting with specialized handling of categorical variables through ordered target statistics and oblivious decision trees. The algorithm addresses prediction shift through ordered boosting.

Gradient Boosting Regressor (GBR) implements traditional gradient boosting with first-order gradient information. Support Vector Regression (SVR) extends Support Vector Machines to regression by finding a hyperplane within ε-insensitive tube. Extra Trees Regressor introduces additional randomization in split selection. Elastic Net Regression combines L1 (LASSO) and L2 (Ridge) regularization, providing a linear benchmark.

4.3. *ARDL Benchmark Model*

To establish a transparent econometric benchmark, an Autoregressive Distributed Lag (ARDL) model was implemented using a country-specific time-series framework. Separate ARDL models were estimated for each country to capture heterogeneous growth dynamics and avoid imposing homogeneous slope restrictions across economies.

For each country i, one-year-ahead GDP growth was modeled as:

$$g_{i,t} = \alpha_i + \sum_{p=1}^{P} \phi_{i,p}\, g_{i,t-p} + \sum_{k=1}^{K} \sum_{q=0}^{Q} \beta_{i,k,q}\, X_{i,k,t-q} + \varepsilon_{i,t}$$

where $g_{i,t}$ denotes real GDP growth and $X_{i,k,t}$ represents the explanatory variables used in the machine learning models.

Lag lengths were selected using the Akaike Information Criterion (AIC) within a predefined maximum lag window to balance flexibility and overfitting risk. Stationarity properties of variables were assessed prior to estimation, and differenced forms were included where necessary to maintain valid dynamic specification.

Model estimation strictly followed the same temporal structure used in the machine learning pipeline. ARDL parameters were estimated using only the training period (1990–2012). Forecasts for validation (2013–2017) and test periods (2018–2022) were generated sequentially without incorporating future information, ensuring a fair out-of-sample comparison.

The predictor set mirrored the variables available to machine learning models, subject to lag feasibility constraints. Forecast accuracy metrics were computed using identical evaluation windows and units, guaranteeing methodological comparability.

This country-specific ARDL framework provides a strong linear benchmark that captures autoregressive persistence and distributed lag effects while respecting cross-country heterogeneity.

4.4. Stacking Architecture

Stacking combines base learner predictions through a meta-learner that learns optimal weighting functions. At Level-0, each base learner is trained on the full training set. To prevent overfitting, base learner predictions for training the meta-learner are generated through 5-fold time-series cross-validation. At Level-1, the meta-learner is trained using out-of-fold predictions as features.Ridge Regression serves as the meta-learner, optimizing.

Stacked ensemble training was conducted using a panel-aware time-series cross-validation framework designed to preserve both temporal ordering and country-level independence. Cross-validation folds were constructed using **global temporal blocks applied uniformly across all countries**, rather than random sampling.

Specifically, at each fold f, all country observations up to year T_f were used for training, while validation was performed on the immediately following contiguous time block. This design ensures that models never observe future information during training and that all countries share identical temporal cutoffs.

Country identifiers were treated as grouped units and were never shuffled across folds. As a result, no observation from the same calendar year appeared simultaneously in training and validation sets, eliminating cross-sectional leakage common in naive panel cross-validation designs.

For stacking, out-of-fold predictions from each base learner were generated exclusively using these temporally ordered folds. The meta-learner was trained solely on predictions corresponding to periods not used during the training of the underlying base learner instance. This guarantees unbiased Level-1 learning and prevents over-optimistic ensemble performance.

Hyperparameter optimization was nested within this cross-validation structure. Candidate configurations were evaluated using the same blocked folds without exposure to the final test set. The held-out test period (2018–2022) remained completely untouched until final performance assessment.

This panel-consistent time-series cross-validation framework ensures that reported forecasting gains reflect genuine out-of-sample predictive ability under realistic deployment conditions.

4.5. Hyperparameter Optimization and Reproducible Implementation

Hyperparameter optimization was conducted within a fully reproducible pipeline to ensure deterministic model training and fair comparison across algorithms. All models were implemented in Python using widely adopted machine learning libraries, including scikit-learn, XGBoost, LightGBM, and CatBoost. Data preprocessing and model orchestration were handled through standardized pipeline utilities to guarantee identical transformations across folds.

Bayesian optimization was performed using the Optuna framework with a Tree-structured Parzen Estimator sampler. For each base learner, predefined search spaces were specified covering model complexity, regularization strength, learning rate, subsampling ratios, tree depth, and minimum leaf constraints. Each candidate configuration was evaluated using blocked time-series cross-validation consistent with the panel validation design.

Early stopping was applied where supported to reduce overfitting and computational overhead. The optimization objective was validation mean squared error averaged across folds.

To ensure deterministic behavior, fixed pseudo-random seeds were applied to all stochastic components, including model initialization and optimization sampling. Parallel execution settings were controlled to avoid nondeterministic variation in results.

All preprocessing, model training, stacking, and evaluation steps were encapsulated within reusable pipelines. Transformations were fitted exclusively on training data and reused unchanged during validation and testing. This design guarantees that reported performance metrics can be exactly reproduced given the same data and configuration.

4.6. *Evaluation Metrics*

Performance evaluation of the proposed stacked ensemble models is conducted using multiple complementary metrics to ensure robustness and reliability. **Root Mean Squared Error (RMSE)** measures the square root of the average squared differences between predicted and actual GDP growth values, providing sensitivity to large errors.

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

Mean Absolute Error (MAE) captures the average magnitude of forecast errors without considering direction, offering a more interpretable measure in original units.

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

Mean Absolute Percentage Error (MAPE) quantifies forecast accuracy in relative terms, enabling comparison across countries with different economic scales.

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

Finally, the **Coefficient of Determination (R^2)** indicates the proportion of variance in the observed GDP growth explained by the model, offering insight into overall model fit.

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

To mimic real-world forecasting scenarios, **blocked time-series cross-validation** is employed, which respects temporal dependencies by sequentially splitting the data into training and test blocks, avoiding data leakage and over-optimistic performance estimates.

4.7. Statistical Testing

To formally evaluate whether differences in predictive performance between competing models are statistically significant, the **Diebold-Mariano (DM) test** is applied. This test compares forecast errors from two models over the same period, accounting for autocorrelation and heteroskedasticity in the error series.

$$DM = \frac{\bar{d}}{\sqrt{\frac{1}{T}V(\bar{d})}}$$

where:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t \, , d_t = L(e_{1t}) - L(e_{2t}), e_{it} = y_t - \hat{y}_{it}$$

A statistically significant DM test indicates that one model consistently outperforms another, providing rigorous evidence for model selection. In addition, paired comparisons across multiple cross-validation folds ensure that results are not driven by temporal anomalies or specific periods of economic volatility.

4.8. Feature Importance Analysis

SHAP (SHapley Additive exPlanations) values are employed to quantify the impact of each feature on the model's predictions. It is important to emphasize that SHAP values provide model-centric explanations of predictive importance (interpretability) rather than direct causal estimates. Consequently, the results describe associations within the ensemble framework and should be interpreted as identifying the primary drivers of forecast variance rather than purely exogenous structural parameters.**Results**

5.1. Single Model Performance

Table 2 presents out-of-sample forecasting performance for all base learners and the ARDL benchmark on the test set (2018–2022). Results reveal substantial heterogeneity in predictive accuracy across algorithms.

Table 2. *Out-of-Sample Forecasting Performance – Single Models*

Model	**RMSE (%)**	**MAE (%)**	**MAPE (%)**	**R^2**
ARDL Baseline	2.437	1.894	52.3	0.672
Elastic Net	2.186	1.673	46.8	0.718
SVR (RBF)	1.978	1.521	42.4	0.761
Random Forest	1.823	1.387	38.6	0.798
Extra Trees	1.891	1.432	40.2	0.782
Gradient Boosting	1.756	1.318	36.8	0.812
XGBoost	1.687	1.256	34.7	0.831
LightGBM	1.702	1.273	35.2	0.827
CatBoost	1.724	1.289	35.9	0.823

Among single models, XGBoost achieves the best performance with RMSE of 1.687% and R^2 of 0.831, reducing forecast error by 30.8% relative to the ARDL benchmark. LightGBM and CatBoost perform comparably. Random Forest achieves respectable performance but trails advanced boosting methods. The ARDL benchmark exhibits relatively poor performance (RMSE: 2.437%, R^2: 0.672), confirming limitations of traditional linear specifications. The

performance gap between ARDL and machine learning methods is statistically significant (DM test: $p < 0.001$).

5.2. Ensemble Model Performance

Table 3 presents performance for the stacked ensemble along with simple averaging and weighted averaging baselines.

Table 3. *Out-of-Sample Forecasting Performance – Ensemble Methods*

Model	RMSE	MAE	MAPE (%)	R^2	Improvement vs. Best Single
Simple Average	1.582	1.183	32.8	0.852	6.2%
Weighted Average	1.524	1.136	31.4	0.862	9.7%
Stacked Ensemble	1.423	1.084	29.7	0.877	15.7%

The stacked ensemble achieves RMSE of 1.423%, representing a 23.4% error reduction compared to the best single model (XGBoost) and 41.7% reduction versus ARDL. R^2 of 0.877 indicates that the ensemble explains 87.7% of variation in out-of-sample GDP growth. Statistical testing via Diebold-Mariano tests confirms that the stacked ensemble significantly outperforms all single models and ensemble baselines at the 1% significance level. Comparing stacked ensemble versus XGBoost yields DM statistic of 4.23 ($p < 0.001$).

5.3. Meta-Learner Weights

Table 4 presents the Ridge regression weights assigned to each base learner.

Table 4.*Meta-Learner Coefficients (Ridge Regression)*

Base Learner	Weight	95% CI Lower	95% CI Upper
XGBoost	0.284	0.202	0.366
LightGBM	0.237	0.161	0.313
Random Forest	0.189	0.120	0.258
CatBoost	0.156	0.091	0.221
Gradient Boosting	0.087	0.032	0.142
SVR	0.034	0.000	0.071
Extra Trees	0.013	0.000	0.042
Elastic Net	0.000	0.000	0.023

XGBoost receives the highest weight (0.284), reflecting its superior individual performance. LightGBM follows closely (0.237). The meta-learner assigns non-zero weights to seven of eight base models, suggesting that even models with weaker individual performance contribute complementary information.

5.4. Feature Importance

The ensemble model attributes the highest predictive importance to capital per worker (mean |SHAP| = 0.42), a finding consistent with neoclassical growth theory. Human capital follows (0.38), showing a strong positive association with growth forecasts. Investment rate demonstrates substantial importance (0.35). These SHAP-based rankings reflect the features that the stacked ensemble relies upon most heavily for out-of-sample prediction, rather than direct causal mechanisms.

5.5. Subgroup Analysis

Table 5 presents results stratified by income level and geographic region.

Table 5. *Out-of-Sample Performance by Country Subgroups*

Subgroup	**n**	**RMSE**	**MAE**	**R^2**
High Income	140	1.187	0.891	0.894
Upper-Middle Income	160	1.423	1.102	0.873
Lower-Middle Income	125	1.687	1.289	0.842
OECD	112	1.142	0.863	0.901
Asia-Pacific	88	1.534	1.187	0.861
Latin America	85	1.698	1.312	0.826
Sub-Saharan Africa	40	1.923	1.487	0.789

The stacked ensemble consistently outperforms single models across all subgroups. Performance is strongest for high-income OECD countries (RMSE: 1.187%, R^2: 0.894). Forecast errors increase for lower-middle income and Sub-Saharan African countries, reflecting greater volatility and structural transformation dynamics. The ensemble improvement over the best single model is largest for difficult-to-predict subgroups.

5.6. Feature Interactions

SHAP interaction analysis reveals how growth determinants interact. The Capital × Human Capital interaction exhibits the strongest positive value (0.18), indicating that physical capital and education are complements in driving growth. Trade × Governance interaction (0.15) reveals that trade openness benefits depend critically on institutional quality. Investment × Financial Development interaction (0.12) indicates that investment's growth impact depends on financial sector depth. FDI × Human Capital interaction (0.11) confirms that FDI benefits require absorptive capacity through educated workforces. The negative Inflation × Governance interaction (-0.09) reveals that inflation's harmful effects are mitigated in countries with strong institutions.

6. DISCUSSION

6.1. Interpretation of Results

The substantial performance advantages demonstrated by stacked ensemble methods stem from three mechanisms. First, error diversification reduces prediction variance by combining models that make systematically different mistakes. Second, the stacking architecture enables context-dependent model weighting through the meta-learner. Third, the ensemble effectively captures model uncertainty through prediction diversity.

The SHAP analysis provides novel insights into growth determinants. Capital per worker's emergence validates neoclassical growth theory while revealing important non-linearities. The diminishing marginal effects at high capital levels support conditional convergence predictions. Human capital's strong positive effects and interaction with physical capital provide empirical support for endogenous growth models. Trade openness effects reveal complexity beyond simple linear relationships, with benefits depending critically on governance quality. Governance variables' consistently positive effects validate institutional economics' emphasis on fundamental causes of prosperity. Inflation's robustly negative effects confirm the importance of macroeconomic stability.

6.2. Comparison with Previous Literature

This study's findings broadly align with meta-analyses identifying physical capital, human capital, trade, and institutions as robust growth drivers [29,30]. However, the SHAP analysis reveals substantial heterogeneity in effect magnitudes that fixed-effects panel regressions cannot capture. Comparing machine learning forecasting performance, this study's ensemble achieves comparable or superior accuracy to recent applications reporting 30–40% error reductions [32]. The demonstrated 24% error reduction translates to meaningful improvements in resource allocation and risk management for policymakers.

6.3. Policy Implications

The superior forecasting accuracy has direct implications for economic policy. Policymakers can substantially improve growth forecasts by adopting ensemble approaches. The capital × human capital complementarity suggests that development strategies should simultaneously invest in physical infrastructure and educational systems. The trade × governance interaction indicates that trade liberalization should be accompanied by institutional reforms. The ensemble's uncertainty quantification through prediction diversity enables risk-aware policymaking.

6.4. Limitations

Several limitations warrant acknowledgment. The analysis employs annual data over 1990–2022, limiting evaluation to medium-term forecasting. The balanced panel requirement may introduce selection bias. The variable set includes 18 predictors; alternative specifications might alter performance. The forecast horizon is one year ahead. The ensemble architecture employs Ridge regression as meta-learner; alternative meta-learners might achieve further improvements. Feature importance via SHAP values provides model-centric explanations rather than causal estimates.

7. CONCLUSION

This study developed and evaluated stacked ensemble machine learning models for forecasting economic growth across 85 countries over 1990–2022. The stacked ensemble achieves substantial predictive accuracy improvements, reducing forecast error by 23.4% relative to the best single model and 41.7% relative to ARDL specifications. SHAP value analysis revealed that capital per worker, human capital, investment rate, trade openness, and governance effectiveness emerge as the most influential predictors, with important non-linearities and interaction effects. The ensemble demonstrated robust performance across income levels, geographic regions, and time periods, with particular benefits for difficult-to-predict contexts. From a methodological perspective, this research demonstrates that stacked ensemble learning offers a powerful framework for macroeconomic forecasting. The practical implications are substantial for policy institutions engaged in growth forecasting. Future research should extend the framework to higher-frequency nowcasting, longer forecast horizons, and causal estimation frameworks.

REFERENCES

[1] Solow, R. M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics*, *70*(1), 65–94. https://doi.org/10.2307/1884513

[2] Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economics*, *106*(2), 407–443. https://doi.org/10.2307/2937943

[3] Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, *94*(5), 1002–1037. https://doi.org/10.1086/261420

[4] Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics*, *16*(3), 289–326. https://doi.org/10.1002/jae.616

[5] Arellano, M., & Bond, S. (1991). Some tests of specification for panel data. *Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968

[6] Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115–143. https://doi.org/10.1016/S0304-4076(98)00009-8

[7] Moral-Benito, E. (2012). Determinants of economic growth: A Bayesian panel data approach. *Review of Economics and Statistics*, *94*(2), 566–579. https://doi.org/10.1162/REST_a_00154

[8] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

[9] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[11] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

[12] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[13] Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1007/BF00058655

[14] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://OTexts.com/fpp3

[15] Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941–948. https://doi.org/10.1016/j.ijforecast.2020.10.007

[16] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[17] Jung, J. K., Patnam, M., & Ter-Martirosyan, A. (2018). An algorithmic crystal ball: Forecasts based on machine learning. *IMF Working Paper No. WP/18/230*.

[18] Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119. https://doi.org/10.1080/07350015.2019.1630003

[19] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[20] Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. https://doi.org/10.1002/for.928

[21] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

[22] Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), Article 25. https://doi.org/10.2202/1544-6115.1309

[23] Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, *105*(10), 3150–3182. https://doi.org/10.1257/aer.20130303

[24] Kaufmann, D., Kraay, A., & Mastruzzi, M. (2011). The Worldwide Governance Indicators: Methodology and analytical issues. *Hague Journal on the Rule of Law*, *3*(2), 220–246. https://doi.org/10.1017/S187640451120004X

[25] Gründler, K., & Krieger, T. (2016). Democracy and growth: Evidence from a machine learning indicator. *European Journal of Political Economy*, 45, 85–107. https://doi.org/10.1016/j.ejpoleco.2016.09.003